DOCUMENT RESUME

ED 445 018 TM 031 555

AUTHOR Taylor, Catherine S.
TITLE Washington Assessment of Student Learning, Grade 4: 1998 Technical Report.
INSTITUTION Washington Office of the State Superintendent of Public Instruction, Olympia.
PUB DATE 1999-06-00
NOTE 130p.
PUB TYPE Guides - Non-Classroom (055)
EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Grade 4; Intermediate Grades; Listening Skills; Mathematics Tests; Reading Tests; *Reliability; *State Programs; *Student Evaluation; *Testing Programs; *Validity; Writing Tests
IDENTIFIERS *Washington

ABSTRACT

This document contains the technical information for the 1998 Washington Assessment of Student Learning (WASL), Grade 4 Assessment for Reading, Mathematics, Listening, and Writing. It documents the technical quality of the assessment, including the evidence for the reliability and validity of test scores. The manual's chapters are: (1) "Overview and Background"; (2) "Item Development and Content Representation"; (3) "Evidence for Validity of Inferences from Test Scores"; (4) "Scoring the WASL Open-Ended Items"; (5) "Standard Setting Procedures"; (6) "Scale Scores"; (7) "Reliability"; and (8) "Description of Performance of Grade 4 Students." Five appendixes contain information on the state's essential academic learning requirements, the test specifications, scoring criteria, and the technical advisory committee. (Contains 57 tables and 3 figures.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 445 018

# Washington Assessment of Student Learning

## Grade 4

## 1998

## Technical Report

Prepared by
Catherine S. Taylor
University of Washington

for
Office of the Superintendent of Public Instruction
P.O. Box 47220
Olympia, Washington 98504-7220

TM031555

BEST COPY AVAILABLE

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

B. J. Patterson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

## TABLE OF CONTENTS


| Part | Title | Pages |
|---|---|---|
| 1 | **Overview and Background** | 1-1 |
| | Washington Assessment System | 1-2 |
| | Criterion-Referenced Testing | 1-6 |
| | Appropriate Use of Test Scores | 1-7 |
| | Description of the Subtests | 1-7 |
| | Estimated Testing Time | 1-10 |
| 2 | **Item Development and Content Representation** | 2-1 |
| | Item and Test Specifications | 2-1 |
| | Content Reviews | 2-4 |
| | Item Tryouts | 2-5 |
| | Scoring and Item Analysis | 2-5 |
| | Rasch Analysis | 2-5 |
| | Traditional Item Analysis | 2-7 |
| | Bias Analysis | 2-8 |
| | Item Selection | 2-10 |
| 3 | **Evidence for Validity of Inferences from Test Scores** | 3-1 |
| | Internal Evidence for Validity of WASL Scores | 3-2 |
| | External Evidence for Validity of WASL Scores | 3-3 |
| | Intercorrelations among WASL and CTBS Scores | 3-3 |
| | Factor Analysis of CTBS and WASL Mathematics Scores | 3-4 |
| | Factor Analysis of CTBS and WASL Reading Scores | 3-7 |
| 4 | **Scoring the WASL Open-Ended Items** | 4-1 |
| | Qualifications of Readers | 4-1 |
| | Range-Finding and Anchor Papers | 4-1 |
| | Training Materials | 4-2 |
| | Rater Consistency (Reliability) | 4-2 |
| | Additional Considerations for Writing | 4-5 |
| 5 | **Standard Setting Procedures** | 5-1 |
| | Reading, Listening, and Mathematics | 5-2 |
| | Writing | 5-4 |

## TABLE OF CONTENTS (Cont.)


| | | |
|---|---|---|
| 6 | **Scale Scores** | 6-1 |
| | Development of Scales Scores on the WASL | 6-1 |
| | Reading and Mathematics | 6-3 |
| | Listening and Writing | 6-4 |
| | Cut Points for Content Strands | 6-4 |
| | Equating | 6-6 |
| | Equating Reading and Mathematics Tests | 6-6 |
| | Equating the Listening Test | 6-7 |
| | Equating the Writing Test | 6-8 |
| | Number Correct Scores to Scale Scores | 6-8 |
| 7 | **Reliability** | |
| | Internal Consistency and Generalizability | 7-1 |
| | Standard Error of Measurement | 7-2 |
| | Interjudge Agreement | 7-3 |
| 8 | **Description of Performance of Grade 4 Students** | 8-1 |
| | Summary Statistics | 8-1 |
| | Percent Meeting Standard | 8-7 |
| | Mean Item Performance and Item-Test Correlations | 8-12 |
| Appendix A | Washington Essential Academic Learning Requirements | |
| Appendix B | Grade 4 Mathematics Test Specifications | |
| Appendix C | Grade 4 Listening and Reading Test Specifications | |
| Appendix D | General Scoring Criteria for WASL Items | |
| Appendix E | National Technical Advisory Committee Members and Washington Assessment Advisory Team Members | |

## TABLE OF TABLES


| Table No. | Title | Page |
|---|---|---|
| Table 1-1 | 1998 Grade 4, Number and Content of Listening Items | 1-8 |
| Table 1-2 | 1998 Grade 4, Number and Content of Reading Items | 1-8 |
| Table 1-3 | 1998 Grade 4, Number and Content of Writing Prompts | 1-9 |
| Table 1-4 | 1998 Grade 4, Number and Content of Mathematics Items | 1-9 |
| Table 1-5 | Estimated Testing Times for Grade 4 WASL | 1-10 |
| Table 2-1 | Grade 4 Reading Subtest: Item distribution by text type, strand, and item type | 2-2 |
| Table 2-2 | Grade 4 Listening Subtest: Item distribution by strand and item type | 2-3 |
| Table 2-3 | Grade 4 Mathematics Subtest: Item distribution by strand and item type | 2-3 |
| Table 2-4 | Responses to Item 3 for Males and Females with Total Test Score of 10 | 2-8 |
| Table 2-5 | Test Development Process for Grade 4 WASL | |
| Table 3-1 | 1998 Grade 4 Correlations Among Content Targets in the Grade 4 WASL | 3-2 |
| Table 3-2 | 1998 Grade 4 Rotated Factor Loadings for Reading and Mathematics Strands | 3-4 |
| Table 3-3 | Grade 4 Multi-Trait/Multi-Method Correlations among 1998 Spring WASL Test Scores and Fall 1997 CTBS Total Scores | 3-4 |
| Table 3-4 | Grade 4 Factor Analysis of 1997 CTBS4 Subtest Scores, 1998 WASL Mathematics strand scores, and 1997 TCS Subtest Scores | 3-6 |
| Table 3-5 | 1998 Grade 4 Correlations Among Latent Factors | 3-6 |
| Table 3-6 | Grade 4 Factor Analysis of 1997 CTBS4 Language Arts Subtest Scores, 1998 WASL Reading strand scores, and 1997 TCS Subtest Scores | 3-8 |
| Table 3-7 | 1998 Grade 4 Correlations Among Latent Factors | 3-8 |
| Table 4-1 | 1998 Grade 4 Correlations and Mean Scores between First and Second Readings of Total Scores for Open-Ended Items by Test | 4-4 |
| Table 4-2 | 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Writing Scores | 4-4 |
| Table 4-3 | 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Listening and Reading Items | 4-4 |

## TABLE OF TABLES (Cont.)


| Table | Title | Page |
|---|---|---|
| Table 4-4 | 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Mathematics Items. | 4-5 |
| Table 5-1 | Number of Standard Setting Judges in each Professional Role | 5-1 |
| Table 5-2 | Example Standard Setting Procedure | 5-3 |
| Table 6-1 | 1998 Grade 4 Listening, Reading, and Writing Number Correct Scores (NCS) to Scale Scores (SS) | 6-9 |
| Table 6-2 | 1998 Grade 4 Mathematics Number Correct Scores (NCS) to Scale Scores (SS) | 6-10 |
| Table 7-3 | 1998 Grade 4 Reliability Estimates (Alpha Coefficient) and Standard Error Of Measurement for Each WASL Test | 7-3 |
| Table 8-1 | 1998 Grade 4 Scale Score Means, Standard Deviations, and Maximum Scale Scores by Subtest | 8-2 |
| Table 8-2 | 1998 Grade 4 Maximum Number Possible, Number Correct Score Means, Standard Deviations (SD) by Strand, and Percent of Students with Strength in Strand | 8-2 |
| Table 8-3 | 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender | 8-3 |
| Table 8-4 | 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group | 8-3 |
| Table 8-5 | 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender | 8-3 |
| Table 3-6 | 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group | 8-3 |
| Table 8-7 | 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender | 8-4 |
| Table 8-8 | 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group | 8-4 |
| Table 8-9 | 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender | 8-4 |
| Table 8-10 | 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group | 8-4 |
| Table 8-11 | 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program | 8-5 |
| Table 8-12 | 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program | 8-5 |

## TABLE OF TABLES (Cont.)


| Table | Title | Page |
|---|---|---|
| Table 8-13 | 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program | 8-6 |
| Table 8-14 | 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program | 8-6 |
| Table 8-15 | 1998 Grade 4 Listening Test: Percent Meeting Standards by Gender | 8-7 |
| Table 8-16 | 1998 Grade 4 Listening Test: Percent Meeting Standards by Ethnic Group | 8-8 |
| Table 8-17 | 1998 Grade 4 Reading Test: Percent Meeting Standards by Gender | 8-8 |
| Table 8-18 | 1998 Grade 4 Reading Test: Percent Meeting Standards by Ethnic Group | 8-8 |
| Table 8-19 | 1998 Grade 4 Writing Test: Percent Meeting Standards by Gender | 8-9 |
| Table 8-20 | 1998 Grade 4 Writing Test: Percent Meeting Standards by Ethnic Group | 8-9 |
| Table 8-21 | 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Gender | 8-9 |
| Table 8-22 | 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Ethnic Group | 8-1 |
| Table 8-23 | 1998 Grade 4 Listening Test: Percent Meeting Standards by Categorical Program | 8-10 |
| Table 8-24 | 1998 Grade 4 Reading Test: Percent Meeting Standards by Categorical Program | 8-11 |
| Table 8-25 | 1998 Grade 4 Writing Test: Percent Meeting Standards by Categorical Program | 8-11 |
| Table 8-26 | 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Categorical Program | 8-12 |
| Table 8-27 | 1998 Grade 4 Listening Test: Number of Points Possible Per Item, Mean Item Performance, and Item-Test Correlation for Each Item | 8-13 |
| Table 8-28 | 1998 Grade 4 Writing Test: Number of Points Possible Per Score-Type, Mean Score, and Score-Total Test Correlation for Each Score | 8-13 |
| Table 8-29 | 1998 Grade 4 Reading Test: Number of Points Possible Per Item, Mean Item Performance, and Item-Test Correlation for Each Item | 8-14 |
| Table 8-30 | 1998 Grade 4 Mathematics Test: Number of Points Possible Per Item, Mean Item Performance, and Item-Test Correlation for Each Item | 8-15 |

## TABLE OF FIGURES


| Figure No. | Title | Page |
|---|---|---|
| Figure 2-1 | Location of examinee $\beta_1$ on two tests with item difficulties $\delta_1$ through $\delta_{10}$ | 2-6 |
| Figure 6-1 | Hypothetical Range of Item Difficulties (theta values) within Mathematics Strands | 6-5 |
| Figure 6-2 | Score Distribution of Students Identified as Below Standard and Score Distribution of Students Identified to Be At or Above Standard: Content, Organization, and Style | 6-5 |

# The Washington Assessment of Student Learning: June 1999

## PURPOSE OF TECHNICAL REPORT

*Standards for Educational and Psychological Testing*[1] (AERA/APA/NCME, 1985) require that test developers and publishers produce a technical manual (p. 35-37, Standards 5.1-5.11). The technical manual must provide overall information documenting the technical quality of the assessment, including evidence for the reliability and validity of test scores. This document contains the technical information for the 1998 *Washington Assessment of Student Learning*: Grade 4 Assessment for Reading, Mathematics, Listening and Writing.

## PART 1

## OVERVIEW

## BACKGROUND FOR THE STATE ASSESSMENT PROGRAM

In 1993, Washington State embarked on the development of a comprehensive school change effort that has as its primary goal the improvement of teaching and learning. Created by the state legislature in 1993, the Commission on Student Learning was charged with three important tasks in support of this school change effort:

- to establish Essential Academic Learning Requirements (EALRs) that describe what all students should know and be able to do in eight content areas--reading, writing, communication, mathematics, science, health/fitness, social studies, and the arts;
- to develop an assessment system to measure student progress at three grade levels towards achieving the EALRs; and
- to recommend an accountability system that recognizes and rewards successful schools and provides support and assistance to less successful schools.

The Commission has achieved its first major task. The EALRs in Reading, Writing, Communications, and Mathematics were first adopted in 1995 and revised in 1997 (See Appendix A). Performance "benchmarks" were also established at three grade levels--elementary (Grade 4), middle (Grade 7), and high school (Grade 10). The EALRs for Science, Social Studies, Health/Fitness, and the Arts were initially adopted in 1996 and also revised in 1997. Performance "benchmarks" for science were also established at three grade levels--elementary (Grade 5), middle (Grade 8), and high school (Grade 10).

---

[1] The latest edition of the Standards for *Educational and Psychological Testing* (AERA/APA/NCME, 1999) were published after the development of this document; however, the report presented in these pages is consistent with the revised standards as well.

The Commission's second major task is to develop an assessment system to determine the extent to which students are achieving the knowledge and skills defined by the EALRs. The assessments for Reading, Writing, Communication, and Mathematics have been developed at Grades 4 and 7 and were both operational as of spring, 1998. The Grade 10 assessment in these same content areas was pilot-tested in spring, 1998 and was operational beginning spring, 1999. Participation in the Grade 4 assessment was mandatory for all public schools beginning spring, 1998. Participation in the Grade 7 and 10 assessments is voluntary until spring, 2001.

Preliminary work is underway to develop middle and high school assessments in Science beginning with pilot assessments in spring, 1999 and operational assessments in spring, 2000. Assessment development work in the other content areas--Social Studies, Health and Fitness, and the Arts--awaits legislative approval and funding.

## WASHINGTON ASSESSMENT SYSTEM

The assessment system has four major components: state-level assessments, classroom-based assessments, professional staff development, and school and system context indicators. These components are described briefly below. Two additional features, the Certificate of Mastery and the Accountability System, are also briefly described.

### State-Level Assessments in Reading, Writing, Listening, and Mathematics

The state-level assessments require students to both select and create answers to demonstrate their knowledge, skills, and understanding in each of the EALRs—from multiple-choice and short-answer items to more extended responses, essays, and problem solving tasks. Student, school, and district scores are reported for the operational assessments. The state-level operational test forms are standardized and "on demand" meaning all students respond to the same items, under the same conditions, and at the same time during the school year.

All of the state-level assessments are untimed; that is, students may have as much time as they reasonably need to complete their work. Guidelines for providing accommodations to students with special needs have been developed to encourage the inclusion of as many students as possible. Special need students include those in special education programs, those with Section 504 plans, English language learners (ESL/bilingual), migrant students, and highly capable students. A broad range of accommodations allows nearly all students access to some or all parts of the assessment (see *Guidelines for Inclusion and Accommodations for Special Populations on State-Level Assessments*).

Classroom teachers and curriculum specialists from across Washington were selected to assist with the development of the items for the state-level assessments. Two content committees were created at each grade level--one for Reading/Writing/Communication and one for Mathematics. Working with content and assessment specialists from the Riverside Publishing Company (one of the Commission's assessment development contractors), these committees defined the test and item specifications consistent with the Washington State Essential Academic Learning Requirements, reviewed all items prior to pilot testing, and

provided final review and approval of all items after pilot testing. A separate "fairness" committee, composed of individuals reflective of Washington's diversity, also reviewed all items for words or content that might be offensive to students or parents, or might disadvantage some students for reasons unrelated to the skill or concept being assessed. (See Part 2 for a more detailed description of this process.)

Literally hundreds of items were developed and pilot-tested to create a "pool" of items. This will allow the creation of new forms of the assessment each year by sampling from the pool. Statistical "equating" procedures are used to maintain the same performance standard from year to year and to provide longitudinal comparisons across years even though different items are used.

The state-level assessments in Reading, Communication, and Mathematics include a mix of multiple-choice, short-answer, and extended-response items. Having a large pool of items provides the opportunity to vary the kinds of items from year to year so that a particular item format (e.g. multiple-choice, short-answer, or extended-response) is not always associated with the same Essential Academic Learning Requirements. (See Part 2 for more detail on the item types)

Following the first operational assessment at each grade level, a standard-setting committee determined the level of performance on the assessments that would be required for students to "meet the standard" on the Essential Academic Learning Requirements. In addition, "progress categories" above and below the standard were established in Reading and Mathematics to show growth over time as well as to give students and parents an indication of how far from the standard in these content areas a student's performance is. School and district performance on the assessments is reported in terms of the percentage of students meeting the standard and in each of the progress categories. (See Part 5 for a complete description of the standard setting process).

An *Example Test* and *Assessment Sampler* for each of the Grade 4, 7, and 10 operational assessments were created for teachers, students, and parents. The *Example Tests* along with the *Assessment Samplers* include samples of the test items, the scoring criteria for the items, and examples of student responses that have been scored. In addition to these materials, an interactive CD-ROM system called NCS Mentor for Washington provides teachers and students with another means to review the Essential Academic Learning Requirements and practice scoring student responses to items like those contained on the operational assessments.

**Classroom-Based Assessment**

There were a number of important reasons for including classroom-based assessment as part of the new assessment system. First, classroom-based assessments help students and their teachers better understand the Essential Academic Learning Requirements and to recognize the characteristics of quality work that define good performance for each content area. Second, classroom-based assessments provide assessment of some of the EALRs for which state-level assessment is not feasible (for example, oral presentations or group discussion). Third, classroom-based assessments offer teachers and students opportunities to

gather evidence of student achievement in ways that best fit the needs and interests of individual students. Fourth, classroom-based assessments help teachers become more effective in gathering valid evidence of student learning related to the Essential Academic Learning Requirements. And finally, good classroom-based assessments can be more sensitive to the developmental needs of students and provide the flexibility necessary to better accommodate the learning styles of children with special needs. In addition to the items that may be on the state-level assessments, classroom-based assessments can provide information from oral interviews and presentations, work products, experiments and projects, or exhibitions of student work collected over a week, a month, or the entire school year.

Classroom-based assessment *Tool Kits* have been developed for the early and middle years to provide teachers with examples of good assessment strategies. The *Tool Kits* include models for paper and pencil tasks, generic checklists of skills and traits, observation assessment strategies, simple rating scales, and generic protocols for oral communications and personal interviews. At the upper grades, classroom-based assessment strategies will also include models for developing and evaluating cross-discipline, performance-based tasks. In addition to the models, the *Tool Kits* also provide content frameworks to assist teachers, at all grade levels, to relate their classroom learning goals and instruction to the Essential Academic Learning Requirements.

**Professional Development**

A third major component of the new assessment system emphasizes the need for ongoing, comprehensive support and professional training for teachers and administrators to improve their understanding of the Essential Academic Learning Requirements, the characteristics of sound assessments, and effective instructional strategies that will help students reach the standards. The Commission on Student Learning established fifteen "Learning and Assessment Centers" across the state. Most are managed through Washington's nine Educational Service Districts with a few managed by school district consortia. These Centers provide professional development and support to assist school and district staff in:

1 linking teaching and curriculum to high academic standards based on the EALRs;

2 learning and applying the principles of good assessment practice;

3 using a variety of assessment techniques and strategies;

4 judging student work by applying explicit scoring criteria;

5 making instructional and curricular decisions based on reliable and valid assessment information; and

6 helping students and parents to understand the EALRs and how students can achieve them.

**Context Indicators**

Context indicators help teachers, parents, and the public understand and interpret student performance in relation to the environment in which teaching and learning occur. Examples of potentially useful indicators include information about faculty experience and training, instructional strategies employed, special programs for students, condition of

facilities and equipment, availability of appropriate instructional materials and technology, relevant characteristics of students and the community, student attendance patterns, grade to grade transition successes, and high school dropout and graduation rates. The purpose for context information is not to explain away or excuse low performance. Rather, context indicators can provide important information to schools, policy-makers, and the public about the conditions that support or inhibit success in helping all students achieve the Essential Academic Learning Requirements.

### Certificate of Mastery

Once the Essential Academic Learning Requirements and new standards are fully in place, graduating seniors will be required to earn a Certificate of Mastery to get a high school diploma. The Certificate will serve as evidence that students have achieved Washington's Essential Academic Learning Requirements by meeting the standards set for the Grade 10 assessments. Preliminary recommendations for implementing the Certificate have been forwarded to the legislature and include the recommendation that initial use should be based only on meeting the standards in Reading, Writing, Communication, and Mathematics. The Certificate as a high school graduation requirement would begin with the graduating class of 2006. Science would be added to the required content areas in 2008. The Commission recommended that meeting the standards in the other content areas be treated as "endorsements" rather than as requirements once those assessments are developed and operational.

### School and District Accountability System

A task force appointed by the Commission has developed preliminary recommendations for a school and district accountability system that will recognize schools who are successful in helping their students achieve the standards on the WASL assessments. Recommendations also address the need for assistance to those schools and districts in which students are not achieving the standards. The task force recommendations were presented to the Commission in June, 1998 (see *Preliminary Accountability System Recommendations for Public Review*).

### Summary

The Commission on Student Learning was committed to developing an instructionally relevant, performance-based assessment system that enhances instruction and student learning. The new assessments are based directly on the EALRs. Therefore, teachers and those who provide pre-service and in-service training to teachers should be thoroughly familiar with the EALRs and the assessments that measure them. Teachers and administrators at all grade levels need to be thinking and talking together about what they must do to prepare students to achieve the EALRs and to demonstrate their achievement on classroom-based and state-level assessments.

## CRITERION-REFERENCED TESTING

The purpose of an achievement test is to determine how well a student has learned important concepts and skills. Test scores are used to make inferences about students' overall performance in a particular domain. In order to decide "how well" a student has done, some external frame of reference is needed. When we compare a student's performance to a desired performance, this is considered a criterion-referenced interpretation. When we compare a student's performance to the performance of other students, this is considered a norm-referenced interpretation.

Criterion-Referenced Tests are intended to provide a measure of the degree to which students have achieved a desired set of learning targets (desired conceptual understandings and skills) that have been identified as appropriate for a given grade or developmental level in school. Careful attention is given to making certain that the items on the test represent only the desired learning targets and that there are sufficient items for each learning target to make dependable statements about students' degree of achievement related to that target. When a standard is set for a criterion-referenced test, examinee scores are compared to the standard in order to draw inferences about whether students have attained the desired level of achievement. Scores on the test are used to make statements like, "this student meets the minimum mathematics requirements for this class," or "this student knows how to apply computational skills to solve a complex word problem."

Norm-Referenced Tests are intended to provide a general measure of some achievement domain. The primary purpose of norm-referenced tests is to make comparisons between students, schools and districts. Careful attention is given to creating items that vary in difficulty so that even the most gifted students may find that some of the items are challenging and even the student who has difficulty in school may respond correctly to some items. Items are included on the test that measure below-grade-level, on-grade-level, and above-grade-level concepts and skills. Items are spread broadly across the domain. While some norm-referenced tests provide objective-level information, items for each objective may represent concepts skills that are not easily learned by most students until later years in school. Examinee scores on a norm-referenced test are compared to the performances of a norm-group (a representative group of students of similar age and grade). Norm groups may be local (other students in a district or state) or national (representative samples of students from throughout the United States). Scores on norm-referenced tests are used to make statements like, "this student is the best student in the class," or "this student knows mathematical concepts better than 75% of the students in the norm group."

To test all of the desired concepts and skills in a domain, testing time would be inordinately long. Well designed state or national achievement tests, whether norm-or criterion-referenced, always include samples from the domain of desired concepts and skills. Therefore, when state or national achievement tests are used, we generalize from a student's performance on the sample of items in the test and estimate how the student would perform in the domain as a whole. To have a broader measure of student achievement in some domain, it is necessary to use more than one assessment. District and classroom assessments are both useful and necessary to supplement information that is derived from state or national achievement tests.

It is possible, sometimes even desirable, to have both norm-referenced and criterion-referenced information about students' performance. The referencing scheme is best determined by the intended use of the test and this is generally determined by how the test is constructed. If tests are being used to make decisions about the success of instruction, the usefulness of an instructional or administrative program, or the degree to which students have attained a set of desired learning targets, then criterion-referenced tests and interpretations are most useful. If the tests are being used to select students for particular programs or compare students, districts, and states, then norm-referenced tests and interpretations are useful. In some cases, both norm-referenced and criterion-referenced interpretations can be made from the same achievement measures. The *Washington Assessment of Student Learning* (WASL) state level assessment is a criterion-referenced test; therefore, student performance should be interpreted in terms of how well students have achieved the Washington state Essential Academic Learning Requirements.

## APPROPRIATE USE OF TEST SCORES

Once tests are administered, WASL performance is reported at the individual, school, and district levels. The information in these reports can be used, along with other assessment information, to help with school and district curriculum planning and classroom instructional decisions. For example, if students in a school are not performing well on the WASL Reading assessment, a careful look at the strand scores (Main Ideas and Details of Fiction; Analysis, Interpretation, and Critique of Fiction, Main Ideas and Details of Non-Fiction; Analysis, Interpretation, and Critique of Non-Fiction) can assist in planning instruction in future years. It may be that students as a whole are successful in comprehending and interpreting literature but are not very successful with informational text. Curriculum planning can center on how to improve materials and instruction related to informational text.

While school and district scores may be useful in curriculum and instructional planning, it is important to exercise extreme caution when interpreting individual reports. The items included on WASL tests are samples from a larger domain. Scores from one test given on a single occasion should never be used to make important decisions about students' placement, the type of instruction they receive, or retention in a given grade level in school. It is important to corroborate individual scores on WASL tests with classroom-based and other local evidence of student learning (e.g., scores from district testing programs). When making decisions about individuals, multiple sources of information should be used and multiple individuals who are familiar with the student's progress and achievement (including parents, teachers, school counselors, school psychologists, specialist teachers, and possibly even the students themselves) should be brought together to make such decisions collaboratively.

## DESCRIPTION OF THE TESTS

The Grade 4 1998 forms of the Washington Assessment of Student Learning measure students' achievement of the Essential Academic Learning Requirements in Reading, Writing, Listening, and Mathematics. The following tables (Tables 1-1 to 1-4) indicate the EALRs measured by each of the four tests, the test "strands", and the number of items per strand in the 1998 test form.

**Table 1-1: 1998 Grade 4, Number and Content of Listening Items**

| Test Strand* | Number of Items |
|---|---|
| Main ideas, details, meaning | 6 |
| Checks for understanding (paraphrasing, questioning, clarifying) | 2 |
| Total No. of Items | 8 |

* Listening EALR 1: The student uses listening and observation skills to gain understanding.

**Table 1-2: 1998 Grade 4, Number and Content of Reading Items**

| Type of Reading Passage | Test Strand | Number of Items |
|---|---|---|
| Fiction (Literary)* | Main ideas, details† | 6 |
| | Analyzes, interprets, and thinks critically† | 6 |
| Non-Fiction (Information or Task Oriented)* | Main ideas, details† | 8 |
| | Analyzes, interprets, and thinks critically† | 10 |
| Total Number of Items | | 30 |

†Reading EALR 2: The student understands the meaning of what is read.

*Reading EALR 3: The student reads different materials for a variety of purposes

**Table 1-3: 1998 Grade 4, Number and Content of Writing Prompts**

| Task | Purposes[1] | Audiences[1] | Process[2] | Number of Prompts | Scores[3] |
|---|---|---|---|---|---|
| Extended Piece | Narrative | Teacher or classmates | • prewrite<br>• first draft<br>• revise<br>• edit<br>• final draft | 1 | • Content, Organization & Style<br>• Writing Mechanics |
| Brief Piece | Persuasive Letter | Business | • prewrite<br>• first draft<br>• revise<br>• edit<br>• final draft | 1 | • Content, Organization & Style<br>• Writing Mechanics |
| Total Number of Prompts | | | | 2 | |

[1] Writing EALR 1: The student writes clearly and effectively (concept & design, style [word choice, sentence fluency, voice], and conventions).

[2] Writing EALR 2: The student writes in a variety of forms for different audiences and purposes.

[3] Writing EALR 3: The student understands and uses the steps of a writing process*

**Table 1-4: 1998 Grade 4, Number and Content of Mathematics Items**

| Process Strand | Concept Strand | Number of Items |
|---|---|---|
| Concepts & Procedures | Number Sense[1] | 6 |
| | Measurement[1] | 5 |
| | Geometric Sense[1] | 4 |
| | Probability and Statistics[1] | 5 |
| | Algebraic Sense[1] | 5 |
| Solves Problems[2] | | 3 |
| Reasons Logically[3] | | 5 |
| Communicates Understanding[4] | | 3 |
| Making Connections[5] | | 4 |
| Total No. of Items | | 40 |

[1] Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.

[2] Mathematics EALR 2: The student solves problems using mathematics.

[3] Mathematics EALR 3: The student uses mathematical reasoning.

[4] Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[5] Mathematics EALR 5: The student makes mathematical connections.

## ESTIMATED TESTING TIME PER SESSION—4TH GRADE - SPRING 1998

The tests in the *Washington Assessment of Student Learning* are not timed. Students should have as much time as they need to work on the tests. Professional judgment should determine when a student is no longer productively engaged. When the majority of students have finished, the few still working may be moved to a new location to finish. Teachers' knowledge of students' work habits or special needs may suggest that some students who work very slowly should be tested separately or grouped with similar students for the entire assessment. For planning purposes, the estimated testing times required for most students are given in Table 1-5.

**Table 1-5: Estimated Testing Times for Grade 4 WASL**

| Session | Subject | Approximate Time[2] |
|---|---|---|
| 1 | Listening | 25 minutes |
| | Reading (Day One) | 50 minutes |
| 2 | Reading (Day Two) | 25 minutes |
| | Writing (Day One) | 50 minutes |
| 3 | Writing (Day Two) | 75 minutes |
| 4 | Mathematics (Day One) with tools | 80 minutes |
| 5 | Mathematics (Day Two) without tools | 80 minutes |

[2] Above times are estimates for actual testing time. Additional time will be required to distribute and collect materials and cover the directions for test-taking. Testing sessions need not follow on consecutive days. Individual sessions should not be split but may be spaced with one or more days in between. However, at Grade 4, Day Two Writing should be scheduled on the next day after Day One Writing.

# PART 2

# TEST DEVELOPMENT AND CONTENT REPRESENTATION

The content of the *Washington Assessment of Student Learning* (WASL) state assessment is derived from the Washington state Essential Academic Learning Requirements (See Appendix A for an overview). These Essential Academic Learning Requirements (EALRs) define, for Washington schools, what students should know and be able to do by the end of grades 4, 7, and 10 in Reading, Writing, Communication, Mathematics, and by the end of grades 5, 8, and 10 in history, geography, economics, civics, science, the arts, health, and fitness. The 1998 WASL subtests measured EALRs for Reading, Writing, Mathematics, and Listening in grades 4 and 7.

## ITEM AND TEST SPECIFICATIONS

The first step in the test development process was to select the "Content Committees" that worked with staff of the Commission on Student Learning (CSL) and the Contractor (Riverside Publishing Company) to develop the actual items, which make up the assessments at each grade level. Each Content Committee was composed of 20 to 25 persons from around the state, most of whom were classroom teachers and curriculum specialists who had teaching experience at or near the grades and in the content areas that were to be assessed (i.e., Reading/Writing/Communication or Mathematics).

The second step in the development process was coming to a common agreement about the meaning and interpretation of the EALRs as well as which ones could be assessed on the state level test. Here it was very important that the Contractor, the Content Committees and the CSL staff were in agreement, in concrete ways, about what students were expected to know and be able to do and how these skills and knowledge would be assessed. In addition, the benchmark indicators were combined in various ways to create testing **targets** for which items would be written (See Appendix B and C).

Next, test specifications were prepared. Test specifications define and describe such details as the kinds and number of items on the assessment, the blueprint or physical layout of the assessment, the amount of time to be devoted to each content area, and the scores to be generated once the test is administered. It was important that the goals of the assessment and the ways in which the results would be used be established at this stage so that the structure of the test would support the intended uses. In addition, the Test Specifications are the basics for developing equivalent test forms in subsequent years as well as creating new items to supplement the item pool. The final Test specifications (See Appendix B and C) document the following topics:

- Purpose of the Assessment
- Strands
- Item Types
- General Considerations of Testing Time and Style
- Test Scoring
- Distribution of Test Items by Item Type

There are three types of items on the *Washington Assessment of Student Learning* (WASL) tests: multiple choice, short answer, and extended response. For each multiple-choice item, students select the one best answer from among three or four choices provided. Each multiple- choice item is worth one point. These items are machine scored.

The other two "open-ended" item types—short answer and extended response—require students to give their own response in words, numbers, or pictures (including graphs or charts). Short-answer items are worth two points (scored 0, 1, or 2) and extended-response items are worth four points (scored 0, 1, 2, 3, or 4). On these items, student responses are assigned partial or full credit based on carefully defined scoring criteria. These items cannot be scored by machine and require hand-scoring by well-trained professional scorers (See Part 4).

In addition to the three item types, students are asked to do two writing assignments (prompts). These prompts may require the students to write a letter requesting information, describe an important event or situation, write a story based on a picture presented, explain a procedure for completing a task or project. etc. Each prompt is worth six points and is hand-scored for content, organization, and style (1, 2, 3, or 4 points) and mechanics and spelling (0, 1, or 2 points).

Tables 2-1 through 2-3 are the test blueprints for item content and item types for the Reading, Listening, and Mathematics subtests of the Grade 4 test.

**Table 2-1: Grade 4 Reading Subtest: Item distribution by text type, strand, and item type**

| Text types/Strands | Number of Reading Selections | Number of Words Per Passage | Number of Multiple-Choice Items | Number of Short Answer Items | Number of Extended Response Items |
|---|---|---|---|---|---|
| **Fictional Text**[1] | 2-3 | up to 750 | 9-12 | 3-5 | 1 |
| Comprehends important ideas and details[2] | | | 3-8 | 1-2 | 0 |
| Analyzes, interprets, and thinks critically[2] | | | 4-8 | 2-4 | 1 |
| **Non-Fiction Text**[1] | 1-2 | up to 750 | 9-12 | 3-5 | 1 |
| Comprehends important ideas and details[2] | | | 3-8 | 1-2 | 0 |
| Analyzes, interprets, and thinks critically[2] | | | 4-8 | 2-4 | 1 |
| Total | 4-5 | up to 1500 | 18-22 | 7-9 | 2 |

[1] Reading EALR 3: The student reads different materials for a variety of purposes

[2] Reading EALR 2: The student understands the meaning of what is read.

**Table 2-2: Grade 4 Listening Subtest: Item distribution by strand and item type**

| Strands | Number of Reading Selections | Number of Words Per Passage | Number of Multiple-Choice Items | Number of Short Answer Items |
|---|---|---|---|---|
| | 1 | up to 200 | 6-8 | 2 |
| Listening for important ideas and details* | | | 6-8 | 0 |
| Paraphrases and summarizes main ideas* | | | 0 | 2 |
| Total | 1 | up to 200 | 6-8 | 2 |

* Communication EALR 1: The student uses listening and observation skills to gain understanding.

**Table 2-3: Grade 4 Mathematics Subtest: Item distribution by strand and item type**

| Strands | Multiple Choice | Short Answer | Extended Response |
|---|---|---|---|
| Number Sense[1] | 3-6 | 1-2 | 0 |
| Measurement Concepts[1] | 3-6 | 1-2 | 0 |
| Geometric Sense[1] | 3-6 | 1-2 | 0 |
| Probability and Statistics Procedures[1] | 3-6 | 1-2 | 0 |
| Algebraic Sense[1] | 3-6 | 1-2 | 0 |
| Solving Problems[2] | 0-2 | 1-2 | 1-2 |
| Reasoning Logically[3] | 0-2 | 1-4 | 0-1 |
| Communicating Understanding[4] | 0-2 | 1-4 | 0-1 |
| Making Connections[5] | 0-2 | 1-4 | 0 |
| Total Number of Items | 24 | 13 | 3 |
| Total Number of Points | 24 | 26 | 12 |

[1] Mathematics EALR 1: The student understands and applies the concepts and procedures of mathematics.

[2] Mathematics EALR 2: The student solves problems using mathematics.

[3] Mathematics EALR 3: The student uses mathematical reasoning.

[4] Mathematics EALR 4: The student communicates knowledge and understanding in mathematical and everyday language.

[5] Mathematics EALR 5: The student makes mathematical connections.

Based on the clarification of the EALRs and the Test Specifications, the next step was to develop Item Specifications. Item specifications provide sufficient detail, including sample items, to direct item writers in the development of appropriate test items for each assessment strand. Separate specifications were produced for the different item types including multiple-choice, short answer and extended response. The Test and Item Specification documents were not only essential for test construction but taken together they will be powerful tools for teachers in developing instructional practices and for administrators in reviewing instructional programs. Test and Item Specifications can be obtained through the web site (www.k12.wa.us) for the Washington State Office of the Superintendent of Public Instruction (OSPI).

## CONTENT REVIEWS

Once the Test and Item Specifications were completed and reviewed by the Content Committees, the Contractor's item writers prepared sample items and scoring criteria to these specifications. The Content Committees task was then to review the items and scoring criteria to assure that the item writers had followed the specifications. As necessary items were revised to ensure that they measured Washington's Essential Academic Learning Requirements both accurately and comprehensively.

When the Content Committees were satisfied that the sample items and scoring criteria were appropriate, the item writers then produced literally hundreds of items to be pilot tested at the selected grade levels. Each test item was coded by content (EALR) area and item type (multiple choice, short answer, extended response) and presented to the Content Committees for final review just as they were to appear on the pilot test forms (including graphics, art work, and location on pages).

When the draft items were completed, the Content Committees reviewed each item focusing on its fit to the Item Specifications, the EALRs, and the appropriateness of item content. For all short answer and extended response items the proposed scoring guidelines (rubrics) were also reviewed. The Committees had three options with each item: approve the item (and scoring guidelines) as presented, recommend changes or actually edit the item (or scoring guidelines) to improve the items "fit" to the EALRs and the Specifications, or eliminate the item from use in the assessment.

In addition to the Content Committees, a separate Fairness Review Committee reviewed each item to identify language or content that might be inappropriate or offensive to students, parents, or communities or items which might contain "stereotypic" or biased references to gender, ethnicity, or culture. As with the Content reviews, The Fairness Review Committee reviewed each item and accepted, edited, or rejected it for use on the pilot assessment.

In order to be included on the pilot assessment, every item was reviewed and approved by both the Content Committees and the Fairness Review Committee. Approved items were to:

- be appropriate measures of the intended content;
- be appropriate in difficulty for the grade level of the examinees;

- have only one correct or best answer for each multiple-choice item;
- have appropriate and complete scoring guidelines for the open response items
- be free from content that might disadvantage some students for reasons unrelated to the concept or skill being tested

## ITEM TRYOUTS

The approved items were then assembled into pilot test forms and administered to carefully-selected, representative samples of students across the state. All schools in the state of Washington were invited to participate in the pilot testing. Eighty five percent of fourth graders took part in the pilots. Test forms were randomly distributed with some effort to ensure that each test form was administered in districts with high populations of ethnic minority students. Each test form was administered to at least 1000 students.

## SCORING AND ITEM ANALYSIS

Following the administration of the pilot assessment, the next step involved scoring the student responses by applying the scoring criteria approved by the Content Committees (See Part 4). A variety of statistical analyses were then employed to determine the effectiveness of the items and to check for item bias that may have been missed by the earlier reviews.

Two methods were used for item analysis. These were: traditional or classical item analysis, which included the item means and item-test correlations for each item, and Rasch analysis, which included the item location and item fit. In addition, bias analysis was conducted using the Mantel-Haenszel bias statistic. Bias analysis investigates whether there is differential item performance for examinees of the same abilities who differ by virtue of gender or ethnicity.

### Rasch Analysis

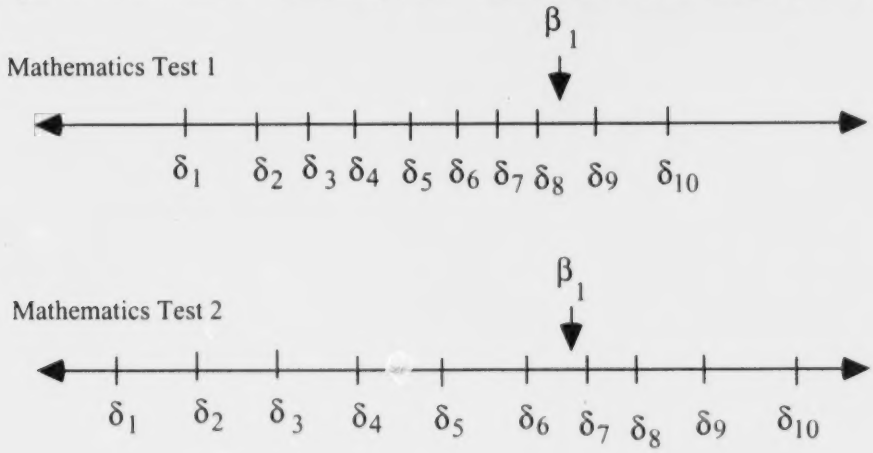
Rasch analysis is an Item Response Theory (IRT) analysis that places all items on a unique continuous scale for each content area. In addition, all examinees in the tryout pool are located on the same underlying scale. The Rasch analysis process separates item difficulty parameters from the abilities of the examinees in the sample that was tested. In this way, item difficulty parameters can be assumed to be the same for groups who are different from the original sample. The basic formula for the Rasch model is:

$$p[x_{vi} = 1 \mid \beta_v, \delta_i] = \frac{\exp(\beta_v - \delta_i)}{1 + \exp(\beta_v - \delta_i)}$$

Where p = the probability of getting an item right given the ability of the examinee ($\beta v$) and the difficulty of the item ($\delta i$).

Working from this formula, item difficulties and examinee abilities can be estimated for a given test. The item difficulty is the point on the ability scale where examinees have a 50/50 chance of getting an item correct. Figure 2-1 shows how examinee ability and item difficulty are placed on ability scales.

**Figure 2-1: Location of examinee $\beta_1$ on two tests with item difficulties $\delta_1$ through $\delta_{10}$**




Because the Rasch model can obtain an equal interval scale independent of item difficulty and person performance, the meaning of test scores can be interpreted in terms of scaled scores rather than number correct scores. For example, in Figure 2-1 (above), the examinee ($\beta_1$) got the first eight items correct on Mathematics Test 1 and the first six items right on Mathematics Test 2. The examinee is the same and her/his mathematics knowledge and skill remains the same; however, the ease or difficulty of the items result in different number correct scores. The Rasch model will indicate the true distance of items from one another across the scale so that examinee test scores reflect the relative distance along the scale rather than the number of items answered correctly. The Rasch model separates item difficulty from examinee ability so that scores of examinees can be interpreted in terms of an underlying ability scale.

For items that have multiple points, a partial credit Rasch model is used to estimate the difficulty (threshold) of each *score* for an item. For example, items with 2 possible points can have two item thresholds: one for the point on the scale (location) at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 0 or 1, and one for the point on the scale at which examinees with abilities equal to that level on the scale have an equal chance of getting a score of 1or 2. The formula for Master's partial credit model (which uses the Rasch dichotomous model as its base) is:

$$\pi_{xvi} = \frac{\exp \sum_{j=0}^{x} (\beta_v - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_v - \delta_{ij})} \qquad x = 0, 1, \ldots m_i$$

Where $\pi$ equals the probability that an examinee with ability $\beta_n$ will get score $x$ on item $i$ and $\delta_{ij}$ is the location of "step" $j$ for item $i$.

Once items and item scores are placed on a scale, items are assessed for "fit" to the Rasch model. The Rasch model assumes there was no guessing on multiple choice items and that, even though the items differ in terms of difficulty (or location on the scale) the items all function equally in discriminating between examinees below and above a given location on the scale. In order to be retained in the item pool, items must measure relevant knowledge and skill, represent desired locations on the ability scale, and fit the Rasch model.

Rasch analyses were conducted independently for each subtest within the Washington Assessment of Student Learning (WASL). The fit of items depends upon whether the items in a scale were all measuring a similar body of knowledge and skill—in other words, whether the scale was unidimensional. Just as height, weight, and body temperature are different dimensions of the human body, so are Reading, Writing, Mathematics, and Listening different dimensions of learning. Therefore, the items and scales for each subtest are examined independently.

In order to place all items across test forms on the same Rasch scale, a subset of items was repeated in adjacent forms. In other words, five items in Form 1 were repeated in Form 2; a different five items in Form 2 were repeated in Form 3; a different five items in Form 3 were repeated in Form 4; a different five items in Form 4 were repeated in Form 5; a different five items in Form 5 were repeated in Form 6; a different five items in Form 6 were repeated in Form 7; a different five items in Form 8 were repeated in Form 1. In this way, Form 1 could be the anchor form and all items could be calibrated back to the item locations for the items in Form 1.

**Traditional Item Analysis**

For multiple-choice items, item means and item-subtest correlations constitute p-values and point-biserials respectively. These are the classical test theory equivalent of item difficulties and item discriminations. The p-value tells the percent of examinees who responded correctly to an item. Its value can range from 0 to 1.0. The point-biserial gives a measure of the relationship between performance on an item and performance on the subtest as a whole and can range from -1.0 to 1.0. Item means indicate, for multiple-point items, the average earned score for examinees in the tryout sample. For 2-point items, item means can range from 0 to 2. For four-point items, item means can range from 0 to 4. Item-subtest correlations, for multiple point items, indicate the relationship between item performance and

subtest performance. Item-subtest correlations can range from -1.0 to 1.0. Item-subtest correlations are computed using the subtest scores relevant to the item.

Unlike the Rasch item data, item means and item-subtest correlations are dependent on the sample of examinees who took the various subtests. If the examinees were exceptionally well schooled in the concepts and skills tested, item means will be fairly high and the items will appear to be easy. If examinees are not well schooled in the concepts and skills tested, item means will be fairly low and items will appear to difficult. If performance on an item does not relate well to performance on the subtest as a whole, item subtest correlations will be low or even negative. Hence both Rasch data and traditional item analysis data are used in item selection.

**Bias Analysis**

The Mantel Haenszel statistic is a chi-square ($\chi^2$) statistic. Examinees are separated into relevant groups based on ethnicity or gender. Examinees in each group are ranked in terms of their total score on the relevant subtest. Examinees in the focal group (e.g., females) are compared with examinees in the reference group (e.g., males) in terms of their performance on individual items. Multiple 2x2 tables are created for each item (one for each total subtest score) indicating, for that score, the number of examinees in each group who got the item right and the number of examinees in each group who got the item wrong. Table 2-4 shows an example 2x2 table for performance on a hypothetical item for males and females with a total subtest score of 10 on a 40 point test. It appears that the item is more difficult for females than it is for males who had a total test score of 10.

**Table 2-4: Responses to Item 3 for Males and Females with Total Test Score of 10**

| Item Number 3 | Number Responding Correctly | Number Responding Incorrectly |
|---|---|---|
| Males (N = 100) | 50 | 50 |
| Females (N = 100) | 30 | 70 |

Examinees with Total Test Score = 10

To complete the Mantel-Haenszel statistic, similar 2x2 tables are created for every test score. A $\chi^2$ statistic is computed for each 2x2 table and the sum of all of the $\chi^2$ statistics across all test scores gives the total bias statistic for a single item. When items have multiple points, a generalized Mantel-Haenszel statistic is computed using all points. Items that demonstrate a high $\Sigma\chi^2$ are flagged for potential bias. Generally, a certain percent of the items in any given pool of items will be flagged for item bias by chance alone. Careful review of items can help to identify whether some characteristic of an item may cause the bias (e.g., the content or language is unfamiliar to girls) or whether the bias data is likely a result of statistical error. For the WASL analyses, the alpha level (error level) was set at .01; therefore, about 1 percent of the items are expected to be flagged for bias by chance alone.

## ITEM SELECTION

Statistical review of items involves examining item means, Rasch item difficulties (locations on the ability scale), and item-test correlations to determine whether items are functioning well. In addition, statistical review requires examining the "fit" of items to the Rasch model. Items that have extremely poor fit to the Rasch model must be revised or removed from the item pool prior to building a final test form. Items that function very poorly (are too easy, too difficult, or have low or negative item-test correlations) must also be revised or removed from the item pool. Finally, items that are flagged for bias against a focal group are examined closely to decide whether they will be removed from the pool. Generally, when item tryouts are conducted, sufficient numbers of items are developed so that revision and new tryouts are not needed. Faulty items can be deleted from the item pool.

After the statistical analyses were completed for the WASL, the Content and Fairness Review Committees reviewed these results and made the final determination about item quality and appropriateness based on the pilot test data. Items were reviewed again for fit to the EALRs; scoring rules were reviewed again for fit to the EALRs and to the demands of the items. In the Fairness Review Committees, bias data were reviewed to determine whether content or language may have resulted in large bias statistics. During these reviews, items were either accepted or rejected for the final pool of items.

Once these reviews were completed the final pool of items was used to develop "operational" test forms. Operational test forms are those that are administered each year to monitor progress of schools and districts in helping students achieve the EALRs. Each operational form is developed by selecting items from the large pool of items tested in the 1996 item tryouts and approved by the Content and Fairness Review Committees. Four criteria are used in item selection for test forms:

1 Item quality
2 Content representation (See Test Specifications)
3 Representation of all gender and ethnic groups (See Test Specifications)
4 Item locations

Item quality is determined by the item means, item-test correlations, bias statistics, Rasch item locations, and fit statistics. Only the best items from the final pool are used in the operational test forms. Test specifications guide item selection to ensure that all relevant strands are represented in each test form as defined in the Test Specifications. Representation of all gender and ethnic groups is reviewed to ensure that Reading and Listening passages and stimulus materials used in the Mathematics and Writing subtests give balanced representations of groups. Finally, because the WASL is intended to be a criterion-referenced test, and because performance standards are established for each subtest, item have been selected to represent a range of locations on the Reading, Mathematics, Writing, and Listening scales. After proficiency scores were established for each subtest in 1997 (See Part 5), item selection for subsequent years has ensured that item locations are similar to those in the initial operational test form in 1997.

Following the administration of the first operational Grade 4 assessment in Spring of 1997, the tests were scored for all participating students. A Standard-Setting Committee (see Part 5) was convened to establish the performance levels appropriate for reporting students' achievement of the EALRs. Based on the standards set by the Committee and approved by the Commission on Student Learning, results for the first Grade 4 operational assessment were reported in September, 1997. Table 2-5 gives the schedule of test development for the Grade 4 WASL.

**Table 2-5: Test Development Process for Grade 4**

| Action | Dates |
|---|---|
| Essential Academic Learning Requirements | March 1995 |
| Test and Item Specifications | Sept - Oct 1995 |
| Item Development | Oct - Dec 1995 |
| Item Review (Content and Fairness) | January 1996 |
| Pilot Testing | May 1996 |
| Item Review (Content and Fairness) | Aug 1996 |
| Item Bank | Sept 1996 |
| Operational Tests Created | Oct - Dec 1996 |
| Published Example Test Assessment Sampler | Feb 1997 |
| First Operational Test Administered | April - May 1997 |
| Standard Setting | June 1997 |
| Score Reports Designed | Sept 1997 |

# PART 3

## EVIDENCE FOR THE VALIDITY OF INFERENCES FROM TEST SCORES

The most important issue in test development is the degree to which the achievement test actually elicits the conceptual understanding and skills that it is supposed to measure. In other words, when one claims that students must use logical reasoning skills to respond to an item, we need evidence that logical reasoning rather than memorization was actually used in the students' responses. Validity is an evaluative judgment about the degree to which the test ***scores*** can be interpreted to mean what test developers claim that they mean. Generally, there are about a half dozen different strategies for obtaining evidence for the validity of test scores (Messick, 1989):

1. We can look at the content of the test in relation to the content of the domain of reference;
2. we can probe the ways in which individuals respond to the items or tasks;
3. we can examine the relationships among responses to the tasks, items, or parts of the test, that is, the internal structure of test responses;
4. we can survey relationships of test scores with other measures and background variables, that is, the test's external structure;
5. we can investigate differences in these test processes and structures over time, across groups and settings, and in response to . . . interventions such as instructional . . . treatment and manipulation of content, task requirements, or motivational conditions;
6. finally, we can trace the social consequences of interpreting and using test scores in particular ways, scrutinizing not only the intended outcomes, but also the unintended side effects. (p. 16)

Validity, then, is a multidimensional construct that resides, not in tests, but in the relationships between any test score and its context (including the instructional practices and the examinee), the knowledge and skills it is to measure, the intended interpretations and uses, and the consequences of its interpretation and use. Messick stated that multiple sources of evidence are needed to investigate the validity of assessments. The following pages provide a description of the evidence available for the validity of scores on the *Washington Assessment of Student Learning* (WASL). This includes: correlations among scores and strands within the WASL; correlations between WASL tests and other achievement tests and measures of ability; and factor analysis studies examining evidence for the construct validity of WASL.

While content representation and item quality are important aspects of tests, they do not ensure the validity of test scores. In order to examine the validity of test scores, it is important to determine whether examinee performance on a test relates to performance on similar measures (external structure) and whether examinees' performance within the set of items on the test is consistent (internal structure). These two types of evidence are considered evidence for the construct validity of test scores. They question whether the test scores elicit the constructs (knowledge and skills) the tests were intended to elicit.

Several studies have been conducted to gather evidence for the construct validity of the WASL Reading, Writing, Listening, and Mathematics tests. The internal structure of the tests have been examined by looking at the intercorrelations among the items and strands assessed by the test. External structure has been examined by looking at correlations among WASL and tests of the *Comprehensive Test of Basic Skills, Fourth Edition* (CTBS), a nationally standardized achievement test. Finally, external structure has been tested through factor analyses of WASL scores, CTBS subtest scores, and subtest scores from the *Test of Cognitive Skills* (TCS), an ability test.

## INTERNAL EVIDENCE FOR THE VALIDITY OF WASL SCORES

### Intercorrelations among WASL Strand Scores

Table 3-1 gives the correlations among the strands within the WASL. As can be seen, scores for Reading strands (Main Ideas and Details in Literature, Analysis, Interpretation, and Critique of Literature, Main Ideas and Details in Informational Text, and Analysis, Interpretation, and Critique of Informational Text) are well correlated (.574 to .679). The Writing Content, Organization, & Style score is moderately correlated with the Writing Mechanics score (.545). Correlations among the Mathematics concepts scores (Number Sense, Measurement, Geometric Sense, Probability and Statistics, and Algebraic Sense) are moderate as would be expected given that these are diverse conceptual areas of Mathematics (.429 to .487). Prior research has shown that students perform differently on mathematical tasks that tap different areas of mathematics (Shavelson, Baxter, & Gao, 1993).

Correlations among the Mathematical process scores (Solves Problems, Reasons Logically, Communicates Understanding, and Making Connections) are also moderate (.397 to .585). The highest correlation is between scores for Reasons Logically and scores for Solves Problems (.585). It is likely that reasoning is an important aspect of problem-solving. Correlations between Mathematics content scores and Mathematics process scores are informative. Scores for Reasons Logically are well correlated with all content strands (.531 to .557) and scores for Makes Connections and Solves Problems are moderately correlated with all content strands (.445 to .475 and .442 to .477 respectively). Since nearly all of the Mathematics items were situated in real world contexts and since all items targeting problem-solving demand the use of mathematical concepts, the moderate correlations make sense. All of the correlations in the mathematics domain reflect moderately positive relationships between different mathematics strands. Correlations between Reading and Mathematics strand scores are also low to moderate (.261 to .565) with most between .40 and .50. Since achievement scores are generally correlated, the strength of the correlations throughout Table 3-1 could be an indication of general achievement.

---

Insert Table 3-1 about here (from end of Part 3)

---

### Factor Analysis of WASL Reading and Mathematics Strand Scores

In order to follow up on these correlations, an exploratory factor analysis was conducted with the Mathematics and Reading strand scores. A principal components analysis was conducted using SPSS. The number of factors was determined using a scree test and finding the solution in which at least 60 percent of the variance was explained. Varimax (orthogonal) rotation was used. The result was a two-factor solution in which 64 percent of the variance was explained. Table 3-2 gives the factor loadings from the rotated component matrix for the two-factor solution. While Reading and Mathematics may be correlated, Reading and Mathematics strands represent separate dimensions of performance on the WASL as a whole.

## EXTERNAL EVIDENCE FOR THE VALIDITY OF WASL SCORES

### Intercorrelations among WASL and CTBS Scores

In order to assess the external validity of WASL scores, a multi-trait/multi-method analysis was conducted using correlations between WASL and CTBS scores from students who were tested in the spring of 1998 with WASL and in the fall of 1997 with CTBS. Table 3-3 gives the correlations among the test scores. There are two methods included in the WASL (multiple choice and open-ended items) and there is one method used in the CTBS (multiple choice). As can be seen, there appears to be a method effect for CTBS and WASL. All of the total scores of CTBS4 are highly intercorrelated (.585 to .830). All of the test scores of WASL are moderately to highly correlated (.347 to .762). This suggests that one of the factors underlying performance on both tests is the way in which the items are structured. Correlations between WASL test scores and CTBS totals are moderately high, suggesting that the constructs measured are all interrelated.

To look more closely at the data, patterns within and between tests were examined. The highest correlation among WASL test scores (see Table 3-3) is between Reading scores and Mathematics scores. The next highest is between Reading scores and Writing scores. The third is between Reading scores and Listening scores. This may suggest that language usage and comprehension are aspects of all four tests. It is evident that WASL Reading scores are moderately to highly correlated with nearly all CTBS scores (.575 to .743) and CTBS Reading Total scores are moderately to highly correlated with all WASL and CTBS4 scores (.375 to .830). The data in Table 3-2, however, show that WASL Reading and Mathematics strand scores reflect different dimensions of achievement. Therefore, although reading may be prerequisite to all achievement, close examination shows that the WASL Mathematics, Writing, and Listening tests are not reading tests.

In looking at the intercorrelations between WASL and CTBS (see Table 3-3), the highest correlation is between WASL Reading scores and CTBS4 Reading Total scores (.743). The next highest correlation is between WASL Reading scores and CTBS4 Language Total scores (.711). The third highest is between WASL Mathematics scores and CTBS4 Mathematics Total scores (.698). All of these correlations provide evidence for the validity of WASL scores beyond the method effect.

**Table 3-2: 1998 Grade 4 Rotated Factor Loadings for Reading and Mathematics Strands for Two Factor Solution**

| Variables | Factor 1 | Factor 2 |
|---|---|---|
| Main Ideas and Details of Fiction | .251 | .824 |
| Analysis, Interpretation, Critique of Fiction | .328 | .805 |
| Main Ideas and Details of Non-fiction Text | .362 | .734 |
| Analysis, Interpretation, Critique of Non-fiction Text | .423 | .728 |
| Number Sense | .688 | .237 |
| Measurement | .605 | .397 |
| Geometric Sense | .641 | .338 |
| Probability and Statistics | .578 | .379 |
| Algebraic Sense | .697 | .237 |
| Solves Problems | .621 | .389 |
| Reasons Logically | .705 | .426 |
| Communicates Understanding | .600 | .188 |
| Makes Connections | .674 | .284 |

Insert Table 3-3 here from end of Part 3

**Factor Analysis of CTBS and WASL Mathematics Scores**

To test whether reading ability, language ability, mathematics ability, or general ability explain WASL Mathematics scores, an exploratory factor analyses was conducted. The data were from fourth grade students who were tested in the fall of 1997 with CTBS and TCS and again in the spring of 1998 with WASL. The variables included in the analysis were WASL scores for Mathematics concepts (Number Sense, Measurement, Geometric Sense, Probability and Statistics, Algebraic Sense) and Mathematics processes (Solves Problems, Reasons Logically, Communicates Understanding, Makes Connections); CTBS subtest scores for Reading Comprehension and Reading Vocabulary (to test the reading hypothesis), Language Mechanics and Language Expression (to test the language hypothesis), and Mathematics Computation and Mathematics Concepts and Applications (to test the mathematics hypothesis); TCS subtest scores for Sequences, Verbal, Analogies, and Memory from the fall of 1997 were included to test the general ability hypothesis.

The abilities that might be expected to predict mathematical performance are sequences and analogies. The sequences subtest measures sequential patterns and our number system is very pattern based. Much of the problem solving and reasoning required in the

WASL problem-solving and reasoning strands requires students to go beyond simple facts and use these facts in new (analogous) situations. If, however, the WASL Mathematics test scores are simply measures of students' verbal ability (a potential alternative explanation of scores), then this would present a problem in understanding the meaning of WASL Mathematics test scores.

A principal components exploratory factor analysis was conducted using SPSS. Three criteria were used to identify the number of factors: eigenvalues greater than 1.0, a scree test, and/or explanation of at least 60 percent of the variance. Given that achievement scores in different domains are generally correlated, an oblique rotation procedure was used (Direct Oblimin).

Table 3-4 gives the factor structure loadings for the analysis of the 1998 Grade 4 data. There were three latent factors and 63 percent of the variance was explained in this analysis. Using a criterion of structure loadings (correlations between the relevant test and the latent factor) greater than or equal to .60 (36% of the variance), the latent factors underlying these test performances appear to be Mathematical Concepts and Applications, General Achievement, and Sequential and Analogical Reasoning. None of the TCS subtest scores loaded with the Mathematics factor. In contrast, Memory and Verbal abilities loaded on the General Achievement factor and Sequence and Analogy subtest scores created a distinct factor. Two subtests from CTBS loaded on the first factor (Mathematics) along with the WASL strand scores: Mathematics Concepts and Applications and Language Expression. This suggests that the 1998 WASL Mathematics test is measuring similar content as CTBS4 Mathematics Concepts and Applications—a point of support for the validity of 1998 WASL Mathematics scores—and that students must use their language skills to express mathematical ideas.

The only WASL Mathematics strand score that loaded with general achievement was Reasons Logically. Table 3-5 shows the correlations among the three latent factors. While there is a fairly strong relationship between Mathematics and General Achievement factors (correlation = .653 or 43% of shared variance), there is only a moderate relationship between General Achievement and Sequential and Analogical Reasoning factors (correlation = .416 or 17% of shared variance). There is a very weak relationship between Mathematics and Sequential and Analogical Reasoning factors (correlation = .354 or 13% of shared variance). These results also suggest that the Sequential and Analogical Reasoning scores are not highly related to 1998 WASL Mathematics test scores, although Memory and Verbal abilities may be modestly related. None of the correlations suggest that these tests could be used interchangeably.

**Table 3-4: Grade 4 Factor Analysis of 1997 CTBS4 Subtest Scores, 1998 WASL Mathematics strand scores, and 1997 TCS Subtest Scores**

| Subtest | Factor 1: Mathematics | Factor 2: General Achievement, Verbal and Memory Abilities | Factor 3: Sequence and Analogical Abilities |
|---|---|---|---|
| WASL Number Sense | .705 | .491 | .306 |
| WASL Measurement | .727 | .516 | .290 |
| WASL Geometric Sense | .694 | .518 | .459 |
| WASL Statistics & Probability | .691 | .511 | .272 |
| WASL Algebraic Sense | .719 | .465 | .305 |
| WASL Solves Problems | .733 | .535 | .282 |
| WASL Reasons Logically | .808 | .607 | .429 |
| WASL Communicates Understanding | .671 | .370 | .005 |
| WASL Makes Connections | .707 | .480 | .438 |
| CTBS4 Reading Vocabulary | .578 | .867 | .291 |
| CTBS4 Reading Comprehension | .581 | .880 | .298 |
| CTBS4 Language Mechanics | .528 | .820 | .372 |
| CTBS4 Language Expression | .605 | .901 | .354 |
| CTBS4 Computation | .552 | .706 | .420 |
| CTBS4 Math Concepts & Application | .663 | .846 | .477 |
| TCS Sequences | .569 | .577 | .759 |
| TCS Analogies | .445 | .505 | .858 |
| TCS Memory | .402 | .702 | .264 |
| TCS Verbal | .523 | .713 | .514 |

**Table 3-5: 1998 Grade 4 Correlations Among Latent Factors**

| Factor | Mathematics | General Achievement, Verbal & Memory Ability | Sequence and Analogical Abilities |
|---|---|---|---|
| Mathematics | 1.00 | .653 | .354 |
| General Achievement | | 1.00 | .416 |
| Sequence and Analogical Abilities | | | 1.00 |

**Factor Analysis of CTBS and WASL Reading Scores**

To test whether reading achievement, language achievement, or general ability explains WASL Reading scores, an exploratory factor analyses was conducted. The data were from fourth grade students who were tested in the fall of 1997 with CTBS and TCS and again in the spring of 1998 with WASL. The variables included in the analysis were WASL scores for Reading strands (Main Ideas and Details of Fiction; Analysis, Interpretation, and Critique of Fiction; Main Ideas and Details of Nonfiction Text; Analysis, Interpretation, and Critique of Nonfiction Text); CTBS4 subtest scores for Reading Comprehension and Reading Vocabulary (to test the reading hypothesis), and Language Mechanics and Language Expression (to test the language hypothesis); TCS subtest scores for Sequences, Verbal, Analogies, and Memory were included to test the general ability hypothesis. Verbal ability might be expected to predict Reading performance. Certainly the capacity to read and understand text is a verbal process. However, if the WASL Reading test scores are simply measures of the kind of thinking relevant to sequences and analogies, then an explanation would be needed for this relationship.

A principal components exploratory factor analysis was conducted using SPSS. Again, three criteria were used to identify the number of factors: eigenvalues greater than 1.0, a scree test, and/or explanation of at least 60 percent of the variance. Again, an oblique rotation procedure was used (Oblimin). Table 3-6 gives the factor structure loadings for the analysis of the 1998 Grade 4 data. There were two latent factors and 67 percent of the variance was explained in this analysis.

Using a criterion of structure loadings (correlations between the relevant test and the latent factor) greater than or equal to .60 (36% of the variance) the factors underlying these test performances suggest that all of the CTBS4 Reading and Language scores as well as the WASL Reading strand scores all load on the same general language arts achievement factor. This suggests that the 1998 WASL Reading test is measuring similar content as CTBS4 Reading Comprehension and, since students must express their understanding of text through writing for some items, Language Expression is required to respond. As with the Mathematics analysis, TCS memory and verbal ability scores loaded on the general language arts achievement factor, suggesting that, for these students, performance on CTBS4 language arts subtests depended on some memory and verbal ability. Three of the TCS subtest scores created a general ability factor: Sequences, Analogies, and Verbal abilities.

Table 3-7 shows the correlations between the latent factors. There is a moderate relationship between the language arts general achievement factor and the general ability factor (correlation = .586 or 34% of shared variance). This suggests that the TCS scores are only moderately related to Language Arts general achievement performance.

The results of the exploratory factor analysis provide good support for the validity of WASL Reading scores. While language usage and vocabulary are required for success on WASL Reading, this should not be surprising given that CTBS4 Reading Total includes Reading Vocabulary and Reading Comprehension is often strongly related to other language arts performance. In contrast, there is no evidence to support a hypothesis that WASL Reading scores are a measure of general ability.

**Table 3-6: Grade 4 Factor Analysis of 1997 CTBS4 Language Arts Subtest Scores, 1998 WASL Reading strand scores, and 1997 TCS Subtest Scores**

| Subtest | Factor 1: Language Arts General Achievement | Factor 2: General Ability |
|---|---|---|
| WASL Main Ideas & Details of Fiction | .770 | .358 |
| WASL Analysis, Interpretation, and Critique of Fiction | .800 | .464 |
| WASL Main Ideas & Details of Non-fiction Text | .757 | .492 |
| WASL Analysis, Interpretation, and Critique of Non-fiction Text | .797 | .527 |
| CTBS4 Reading Vocabulary | .836 | .479 |
| CTBS4 Reading Comprehension | .872 | .482 |
| CTBS4 Language Expression | .878 | .542 |
| CTBS4 Language Mechanics | .754 | .543 |
| TCS Sequences | .566 | .860 |
| TCS Analogies | .503 | .892 |
| TCS Memory | .678 | .379 |
| TCS Verbal | .702 | .637 |

**Table 3-7: Grade 4 Correlations Among Latent Factors**

| Factor | General Language Arts Achievement | General Ability |
|---|---|---|
| General Language Arts Achievement | 1.00 | .586 |
| General Ability | | 1.00 |

**References**

Shavelson, R. J., Baxter, G. P., Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215-232.

# PART 4

# SCORING THE WASL OPEN-ENDED ITEMS

During item development, scoring criteria for each open-ended item on the *Washington Assessment of Student Learning* (WASL) were written. Appendix D provides the general scoring criteria that served as the guides for the item specific scoring criteria for Reading, Mathematics, and Listening items. Appendix D also provides an example of how the general scoring criteria for a mathematics item was made specific to the requirements of the task. During item reviews, the scoring criteria were reviewed along with item directions. A central aspect of the validity and reliability of test scores is the degree to which scoring criteria are related to the appropriate learning targets (Essential Academic Learning Requirements) and whether they are applied faithfully during scoring sessions. Appendix D also provides the scoring criteria for all student writing samples. The following procedures were used to score the WASL items and apply to all content areas that include open-ended questions calling for student constructed responses. These procedures were used for the full pool of items that were pilot tested as well as for the 1998 operational tests.

## QUALIFICATIONS OF READERS

Highly-qualified, experienced readers (scorers) were essential to achieving and maintaining consistency and reliability when scoring student-constructed (open-ended) responses. Readers selected for the Washington Assessment of Student Learning were required to have the following qualifications:

- A minimum of a bachelor's degree in an appropriate academic discipline (such as English, English Education, Math, Math Education, or related fields);
- Demonstrable ability in performance assessment scoring;
- Teaching experience, especially at the elementary or secondary level, was preferred.

Team and table leaders, responsible for supervising small groups of readers, were selected on the basis of demonstrated expertise in all facets of the scoring process, including strong organizational abilities, leadership, and interpersonal communication skills.

## RANGE-FINDING AND ANCHOR PAPERS

The thoughtful selection of papers for range-finding and the subsequent compilation of anchor papers and other training materials were the essential first steps to ensure that scoring was conducted consistently, reliably, and equitably.

In the range-finding process, performance assessment and curriculum specialists working with team and table leaders and teachers from Washington all became thoroughly familiar with and reached consensus on the scoring criteria (rubrics) approved by the Content Committees for each open-ended item. These range-finding teams began work with random selections of student responses for each item. They reviewed these responses, selected an appropriate range of responses, and placed them into packets, numbered for easy reference. The packets of responses were read independently by members of a team of the most

experienced readers. Following these independent readings and tentative ratings of the papers, the total range finding group worked together to discuss both the common and divergent scores. From this work, they assembled tentative sets of example responses for each prompt.

The primary task of the range-finding committee then was the identification of anchor papers—exemplars that clearly and unambiguously represented the solid center of a score point as described in the scoring criteria. Those exemplary anchor papers formed the basis not only of reader training, but of subsequent range-finding discussions, as well.

Discussion was ongoing with the goal of identifying a sufficient pool of additional student responses for which consensus scores could be achieved and which illustrated the full range of student performance in response to the prompt or item. This pool of responses included borderline responses—ones which appeared to be between rather than clearly within a score level and which therefore represented a decision-making problem that readers (with training) would need to resolve.

## TRAINING MATERIALS

Following the range-finding sessions, the performance assessment specialists and team leaders finalized the anchor sets and other training materials, as identified in the range-finding meetings. The final anchor papers were chosen for their clarity in exemplifying the criteria defined in the scoring rubrics.

The anchor set for each 4-point question consisted of a minimum of thirteen papers, three examples of each of the four score points and one example of a non-scorable paper. The anchor set for each 2-point question consisted of a minimum of seven papers, three examples of each of each score point and one example of a non-scorable paper. Score point exemplars consisted of one low, one solid mid-range, and one high example at each score point.

Additional training and qualifying sets of responses were selected to be used in reader training. One training set consisted of responses that were clear-cut examples of each score point; the second set consisted of responses closer to the borderline between two score points. The training sets gave readers an introduction to the variety of responses they would encounter while scoring, as well as allowing them to develop their decision-making capability for scoring responses that did not fall clearly into one of the scoring levels. Calibration/validity papers to be circulated during scoring were also identified at this time, as were reader qualifying sets.

## RATER CONSISTENCY (RELIABILITY)

Reader training for each prompt was led by performance assessment specialists and team leaders. The primary purpose of the training was to help the readers understand the decisions made by the range-finding committee. Also, training helped readers internalize the scoring rubrics, so that they might effectively and consistently apply them.

Reader training sessions included an introduction to the assessment itself. In addition, readers were informed of the parameters or context within which the students' performance was elicited. This gave readers a better understanding of what types of responses could be expected, given such parameters as grade level, instruction or time limitations. Readers next received a description of the scoring criteria that applied to the responses for each item.

The scoring criteria were always presented in conjunction with the anchor papers. After presentation and discussion of the anchor papers, each reader was given a training set consisting of ten papers. The readers scored the papers independently. When all readers had scored the training set, their preliminary scores were collected for reference.

Group discussion of the scores assigned was the next step, allowing the readers to raise questions about the application of the scoring rubric and giving them a context for those questions. The purpose of the discussion among the readers in training was to establish a consensus to ensure consistency of scores between readers. Even after readers had qualified for the scoring, training continued throughout the scoring of all responses to maintain high inter- and intra-reader reliability. Therefore, training was a continuous process and readers were consistently given feedback as they scored.

Frequent reliability checks were used to closely monitor the consistency of each reader's performance over time. The primary method of monitoring a reader's performance was by a process called "back-reading". In back-reading, each table leader reread and checked scores on an average of five to ten percent of each reader's work each day, with a higher percentage early in the scoring. If a reader was consistently assigning scores other than those the table leader would assign, the team leader and performance assessment specialist, together, retrained that reader, using the original anchor papers and training materials. This continuous, on-the-spot checking provided an effective guard against reader "drift," (beginning to score higher or lower than the anchor paper scores). Readers were replaced if they were unable to score consistently with the rubric and the anchor papers after significant training.

Tables 4-1 through 4-4 give the rater agreement information for the open-ended items in the 1998 WASL. Two types of rater agreement were calculated: score agreement for individual items and score agreement across the total score for the open-ended item set for each content area. For total score agreement on the open-ended items, the correlations were quite high (.96 to .98) within each content area. For item-by-item interjudge agreement in Reading and Listening, the range of exact agreement is 76 to 97 percent and the range of exact and adjacent agreement was 99 to 100 percent. For item-by-item interjudge agreement in Mathematics, the range of exact agreement was 79 to 96 percent and the range of exact and adjacent agreement was 95 to 99 percent. For item-by-item interjudge agreement in Writing, the range of exact agreement was 86 to 90 percent; exact and adjacent agreement was approximately 100 percent.

**Table 4-1: 1998 Grade 4 Correlations between and Means of Total Scores of First and Second Readings for Open-Ended Items by Test.**

| Test | Correlation | Mean First Reading | Mean Second Reading |
|---|---|---|---|
| Listening & Reading | .98 | 13.52 | 13.48 |
| Writing | .96 | 7.37 | 7.37 |
| Mathematics | .98 | 17.83 | 17.81 |

**Table 4-2: 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Writing Scores.**

| Score | Points Possible | Exact Score Match | Adjacent Scores | Discrepant by Two Points | Discrepant by Three Points | Discrepant by Four Points | Discrepant by Five Points |
|---|---|---|---|---|---|---|---|
| 1 | 4 | 7578 | 1197 | 10 | | | |
| 2 | 2 | 7912 | 873 | | | | |
| 3 | 4 | 7670 | 1101 | 9 | 3 | 2 | |
| 4 | 2 | 7919 | 863 | 3 | | | |

**Table 4-3: 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Listening and Reading Items.**

| Item | Points Possible | Exact Score Match | Adjacent Scores | Discrepant by Two Points | Discrepant by Three Points | Discrepant by Four Points |
|---|---|---|---|---|---|---|
| 1* | 2 | 6578 | 1211 | 6 | | |
| 6* | 2 | 6974 | 811 | 10 | | |
| 5 | 2 | 6455 | 1290 | 50 | | |
| 6 | 2 | 7278 | 513 | 4 | | |
| 14 | 2 | 6899 | 894 | 2 | | |
| 16 | 2 | 6845 | 916 | 34 | | |
| 17 | 4 | 5928 | 1794 | 66 | 6 | 1 |
| 19 | 2 | 7408 | 384 | 3 | | |
| 22 | 2 | 7430 | 360 | 5 | | |
| 23 | 4 | 5956 | 1731 | 94 | 10 | 4 |
| 25 | 2 | 7529 | 258 | 8 | | |

* Listening items

**Table 4-4: 1998 Grade 4 Frequencies of Exact Score Matches, Adjacent Scores, and Discrepant Scores for Mathematics Items.**

| Item | Points Possible | Exact Score Match | Adjacent Scores | Discrepant by Two Points | Discrepant by Three Points | Discrepant by Four Points |
|---|---|---|---|---|---|---|
| 3 | 2 | 7005 | 766 | 10 | | |
| 5 | 2 | 7332 | 396 | 53 | | |
| 8 | 2 | 7398 | 365 | 18 | | |
| 10 | 4 | 6180 | 1214 | 275 | 99 | 13 |
| 14 | 2 | 6633 | 1112 | 36 | | |
| 18 | 4 | 6652 | 946 | 168 | 14 | 1 |
| 20 | 2 | 7273 | 425 | 83 | | |
| 23 | 2 | 7087 | 656 | 38 | | |
| 25 | 2 | 7260 | 454 | 67 | | |
| 26 | 2 | 6616 | 1059 | 106 | | |
| 29 | 2 | 6393 | 1338 | 50 | | |
| 32 | 2 | 7272 | 458 | 51 | | |
| 34 | 2 | 6755 | 1021 | 5 | | |
| 35 | 2 | 6307 | 1409 | 65 | | |
| 37 | 4 | 6969 | 702 | 88 | 18 | 4 |
| 39 | 2 | 7463 | 293 | 25 | | |

### Additional Considerations For Scoring Writing

Although the training for scoring writing is the same as described above, various approaches can be used in scoring Writing. For the WASL, a "focused holistic" approach was selected. Focused holistic scoring, or general impression scoring, assesses relative writing fluency and measures the degree to which a writer has connected to the reader of a paper. When a paper is scored holistically, a reader considers the overall effectiveness of the piece of writing and assigns a score that reflects the reader's impression of the paper's overall quality. In a focused holistic approach, the reader also takes into account all of the elements that make up a successful piece of writing, for example content, organization, style, and mechanics. In the WASL Writing Test, Content, Organization, and Style are scored together on a 4-point scale and Writing Mechanics are scored on a 2-point scale. These two scores are combined to provide a maximum of 6 points on any one piece of writing.

# PART 5

# STANDARD SETTING PROCEDURES

Standard setting for the Grade 4 *Washington Assessment of Student Learning* (WASL) was conducted in the summer of 1997. Because all of the items in the WASL item pool are on the same underlying Rasch scale (see Part 2), these standards can be held consistent across different test forms, making it possible to monitor student achievement over time with a fixed performance standard in each content area.

Standard setting committees were composed of teachers, curriculum specialists in the relevant subject area, school administrators, parents, and community members (Table 5-1). All standard setting committee members had direct experience with fourth graders or with the curriculum materials relevant for fourth graders.

**Table 5-1: Number of Standard Setting Judges in each Professional Role.**

| Professional Role | Number of Judges |
|---|---|
| Elementary Teachers | 25 |
| Specialist Teachers | 2 |
| School Administrator | 4 |
| Parent | 5 |
| Community Representative | 3 |
| Total | 39 |

Setting standards for student performance on the WASL was essentially a systematic, judgmental process aimed at establishing a consensus, among knowledgeable people, about what fourth grade students should know and be able to do. Washington's Essential Academic Learning Requirements (EALRs) have described the expected content in Reading, Writing, Communications, and Mathematics for Washington's public schools (See Appendix A). The new assessments have defined, in performance terms, some of the important knowledge, skills, and abilities fourth grade students should demonstrate in relation to the EALRs. The purpose of the standard-setting process was to establish the level of performance expected of fourth grade students who are judged as meeting the standards in Listening, Reading, Writing, and Mathematics. The emphasis for the judges, in the standard setting process, was on what students should know and be able to do near the end of Grade 4.

Performance standards on the Grade 4 assessment were determined by the standard setting procedure described below. This procedure is particularly well adapted to setting standards on assessments with mixed item types (that is, multiple-choice, short-answer, and extended response formats) as used on WASL. The procedure used in Washington state has been applied successfully in other large-scale assessment programs and was reviewed and approved by the National Technical Advisory Committee (see Appendix E) for the Commission on Student Learning—a committee composed of nationally recognized measurement professionals.

## READING, LISTENING, AND MATHEMATICS

Implementation of the standard setting process required that the judges first take the operational test just as the students experienced it. The judges also reviewed scoring guides for the constructed-response (short-answer and open-ended) items and examples of student responses anchoring each item's score points.

Next, each standard setting judge received a complete set of the items ordered by difficulty from easiest to hardest, rather than in the order they appeared in the students' test booklets. Multiple-choice items appeared only once in the ordered booklet. Two-and four-point items appeared two or four times, according to the difficulty of achieving each score point. Data from the spring 1997 operational assessment was used to establish item difficulties. The first item in the judges' ordered booklets was the easiest item on the test, that is, the one the highest number of students answered correctly. The last item in the judges' ordered booklets was the hardest item on the test, that is, the one the fewest number of students answered correctly. Although the judges knew the items were ordered from easiest to most difficult, they did not know how students actually performed on the items—that is, how many students answered item 1 correctly, item 2 correctly, and so forth.

In small groups, the judges examined the items in the ordered booklet one at a time, starting with the first (easiest) item in the booklet, and moving to the second easiest item, and so on, until all items (and their scoring rubrics) were examined. As judges examined each item, they were asked to consider:

- What is each item measuring?
- What makes each item more difficult than the items that precede it?

Judges proceeded through the ordered item booklets and trained table leaders encouraged them to observe the increase in the complexity of the items and note the increase in knowledge, skills, and abilities required to answer the items.

At the conclusion of this first review of the ordered booklets, judges were asked to make an individual decision about where to place a "flag" at "meets standard". Each flag was placed in the ordered item booklet according to the individual judge's expectation of what students who are performing at standard should know and be able to do. For example, each judge placed his or her "meets standard" flag at a location in the booklet such that if a student is able to respond correctly to the items that precede the flag (with at least 2/3 likelihood of success), then the student has demonstrated sufficient knowledge, skills, and abilities to infer that the student is performing at the standard. For multiple-choice items this means the student who "meets standard" should be likely to know the correct response. For short answer- or extended response-items (with multiple score points), this means the student who "meets standard" should be likely to achieve at least that score point.

For the Reading and Mathematics tests, judges were asked to insert two additional flags: one at "exceeds standard" and one between "near standard" (partially proficient) and "low" (minimal). In this way, progress toward or beyond standards could also be identified. These additional flags were not set for the Listening test because there were not a sufficient number of points on each test to warrant such a fine distinction of performance levels.

Because not all judges set their flags in the same locations, the next step involved each judge sharing and discussing the locations at which his or her flag(s) were placed. When one judge placed a flag for "meets standard" farther along in the ordered booklet than another judge, it implied that the first judge expected students who meet the standard to demonstrate a higher level of achievement on the test. The difference in their individual expectations was reflected by the content and difficulty of the items between their flags.

For example, if Judge 1 placed a flag after item 30 and Judge 2 placed a flag after item 40, then these two judges disagreed on items 31-40. We know this because Judge 1, who placed a flag after item 30 was indicating that students who can correctly respond to the content in items 1-30 (with at least 2/3 likelihood) have demonstrated abilities sufficient to infer they have met the standard. Judge 2 (who placed the flag after item 40) did not agree, and was indicating that students have not demonstrated sufficient skills until they can handle more difficult content, that is, items 31-40.

Judges next discussed in small groups these differences in expectations as indicated by their different flag placements. Each group was provided with three lists indicating each judge's three flag locations for Reading and Mathematics. Beginning with the judges' placements of the "meets standard" flags, each judge was asked to note the location of every other judge's flag placement. Suppose the results in Table 5-2 occurred from the first round of standard setting.

**Table 5-2: Example of Standard Setting Procedure**

| Judge Number | Meets Standard Flag Placed After: |
|---|---|
| 1 | item 30 |
| 2 | item 34 |
| 3 | item 29 |
| 4 | item 33 |
| 5 | item 36 |
| 6 | item 39 |
| 7 | item 33 |

Judges next would be asked to place a flag in their own ordered booklets after items 29, 30, 33, 34, 36, and 39. Now all judges could see the different expectations for student performance that "meet standard." In this example, judges would next discuss their differences, focusing on the items between 30 and 39 and discuss what these items ask of students' knowledge, skills, and abilities and whether students who meet the standard should be expected to respond correctly to these items. The discussion would consider the items one at a time beginning with item 30 and continuing up through item 39. When productive

discussion of these items was completed, judges would then be asked to reevaluate their own initial flag locations in light of the small group discussion. Judges may decide to agree on a common flag placement during this round. That is, rather than requiring the calculation of the small group's average to determine the group's flag placement, the judges may agree to compromise and reach a consensus.

In the standard-setting for Reading and Mathematics, after judges had made their second round flag placements for "meets standard", the process was repeated for the other two cut-points—the below standard and the above standard locations.

Round 3 consisted of bringing the small groups back together as a large group to share and discuss each small group's flag placements. In the large group each judge placed a flag in his/her own ordered item booklet where each small group had made its flag placements. Large group discussion now focused on the items between the first and last flags for each performance level. Following the large group discussion, judges were asked to make a new (or reconfirm their former) flag placements.

Round 4 consisted of sharing with the large group the Round 3 small group results. Individual judges were then asked to make their final post-it flag placements, which were then compiled to establish the final standard and other performance levels for each content area.

## WRITING

Writing was handled in a slightly different manner than for Reading, Listening and Mathematics. There were two prompts (writing tasks). Each was scored for Content, Organization, and Style (1-4 points) and Mechanics (0-2 points). The scores from both prompts were combined (a possible range of 2-12 points) and the standard was set on the combined scores. To keep the standard-setting process for Writing as parallel with the other content areas as possible, the following standard-setting procedure was used:

1. Example responses were selected (both prompts together from the same student) that represented each of the possible combined score points 2-12 using a minimum of 3 students' responses for each possible score point.
2. These sets of combined student responses were ordered from lowest combined score (2) to highest combined score (12).
3. Judges were asked to proceed individually through all the example response sets (a minimum of 33) from lowest to highest and indicate the point at which the papers began to represent work "at the standard" and prior sets of papers represented work that was "less than the standard."
4. Next judges shared their individual judgments in their small groups and discussed the characteristics of the papers just above and just below their cut-points (post-it flags).
5. The small group's placements were shared and discussed in the larger group.
6. Finally judges reconsidered their post-it flags in light of the discussions and worked toward a consensus as to where the standard for Writing should be set.

## SUMMARY

These processes ensured that the standards set for proficiency on the WASL tests would have careful scrutiny from a broad range of constituents of education. The judges had significant input from their peers and sufficient opportunities for discussion about their diverse opinions on standards.

# PART 6

# SCALE SCORES

All scaling for the grade 4 Washington Assessment of Student Learning (WASL) was done using the same item data and calibrations used in the standard setting. Because the Mathematics and Reading tests have four levels for student performance versus two levels for Listening and Writing, two different procedures were used to develop the scale scores. All four of the tests have a scale score of 400 representing the standard, but for Reading and Mathematics, the cut score for level two was set to equal 375 whereas in Listening and Writing an adjustment to the standard deviations was made to produce the scale scores. The following sections give details pertaining to the actual procedures used

## DEVELOPMENT OF SCALE SCORES ON THE WASHINGTON ASSESSMENT OF STUDENT LEARNING

Scores on the WASL are reported as scale scores (see Tables 6-1 and 6-2 on Pages 9 and 10 of Part 6 for 1998 Grade 4 number correct to scale scores conversions for each test). As described in Part 2, the Rasch model and Master's (1982) extension of the Rasch model to multiple point items (the partial credit model) result in an equal interval scale (much like a ruler that is marked in inches or centimeters) for each test on which items and student scores can be reported. The partial credit model allows for the inclusion of open-ended items where the maximum points possible are greater than one. Calibrating a test with Master's partial credit model produces estimated item parameters for an item's difficulty and the difficulty of its various score points (or steps). The possible scale score range for the WASL across the four test scales is 150 to 600 given all of the items in the item pool. This range is sufficient to describe levels of performance from the lowest possible earned scale score to the highest possible earned scale score *across all content areas tested and across different test forms*. The actual range of scale scores each year and in each content area will differ. For example, for the Grade 4 Mathematics test, the actual range of scale scores in 1998 is 195 to 552.

The Rasch model is an item response theory (IRT) model. IRT models can generate three parameters for items: item difficulties, item discriminations, and guess levels (the probability that low achieving examinees can guess correctly on multiple-choice items). The Rasch and PCM models also generate theta ($\theta$) for each examinee. Because Rasch models treat all items as equally discriminating and assume that there is no guessing, there are no item discrimination and guessing parameters calculated. This means that, unlike more complicated scoring models, there is a one to one relationship between the number correct score on a test and the $\theta$ score on the test.

Once $\theta$ scores are generated, it is general practice to convert $\theta$ to a positive, whole number scale through a linear conversion procedure. The resulting numbers on the whole number scale are easy to use for computations when generating district, school, or building averages.

Because the scaled scores are on an equal interval scale, it is possible to compare score performance at different points on the scale. Much like a yard-stick, differences are

constant at different measurement points. For example, a difference of 2 inches between 12 and 14 inches is the same differences as a difference of 2 inches between 30 and 32 inches. Two inches is two inches. Similarly, for equal interval achievement scales, a difference of 40 scaled score points between 360 and 380 means the same difference in achievement as a difference of 400 and 420, except that the difference is in degree of achievement rather than length.

The major limitation of scaled scores is that they are not well suited to making score interpretations beyond "how much more" and "how much less". Administrators, parents, and students ask, "What score is good enough? How do we compare with other schools like ours? Is a 40 point difference between our school and another school a meaningful difference?" For this reason, scale scores are usually interpreted by using performance standards or converting them to percentile ranks.

Based on the content of the WASL, committees set the performance standards for each test (Reading, Writing, Listening, and Mathematics) that would represent acceptable performance for a well taught, hard working fourth grade student (see Part 4). In Reading and Mathematics, the standard setting committees also identified two "below standard" and one "above standard" performance levels[1]. Because the Listening and Writing tests were relatively short, only two performance levels were established - "meets standard" and "does not meet standard."

The standard setting (described in Part 4) allowed the standard setting committees to identify the $\theta$ values associated with each cut-score (i.e., in Reading and Mathematics, the cut between "substantially below standard" and "approaches standard", between approaches standard and "meets standard", and finally between "meets standard" and "exceeds standard"; in Writing and Listening, the cut between "does not meet standard" and "meets standard"). It was these $\theta$ values that formed the basics for the scaling procedure. In order to maintain the linear scale defined by the raw score to $\theta$ relationship, any two points on the $\theta$ scale can be

---

[1] The following are the general descriptions of the performance levels established for the Washington Assessment of Student Learning:

Level 4 -- Above Standard: This level represents superior performance, notably above that required for meeting the standard at grade 4.

Level 3 -- MEETS STANDARD: This level represents solid academic performance for grade 4. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 -- Below Standard: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 4.

Level 1 -- Well Below Standard: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at grade 4.

*In all content areas, the standard (Level 3) reflects what a well taught, hard working student should know and be able to do.*

fixed to scale scores and the resulting transformation will remain linear. That is what was done here.

### Reading and Mathematics

Following the standard setting process, a linear conversion was used to transform the θ (logistic ability) scores (from the Rasch and partial-credit model analyses) to a whole number scale. For all tests, the θ score identified as "meets standard" was converted to a WASL scale score of 400. For Reading and Mathematics, the θ score identified as "below standard level 2" was converted to a Washington scale score of 375. The rest of the θ scores were converted to the whole number scale using the linear conversion equations for each test that produced these two scale score points. Only two points can be set in a linear transformation and all other points must be derived from the conversion formula. Therefore, the "above standard" scale score for Reading was set at 421 and the "above standard" scale score for Mathematics was set at 422.

The general formula for a linear equation converting θ to a scaled score is:

$$\theta a + b = \text{scaled score} \qquad (6\text{-}1)$$

Where **a** is a distribution variable for the whole number scaled scores and **b** is a location on the whole number scale.

To obtain the linear formula necessary to translate from the θ scale to the whole number scale for Reading and Mathematics, the scaled score cut points for "meets standard" (400) and approaches standard (375) are plugged into the above formula and, through simultaneous solution of two equations, one can solve for **a** and **b.**

For math, the point on the θ scale where the standard setting committee decided that students had "met standard" was .6815 and the point on the θ scale where the standard setting committee decided that students were "approaching standard" was .021. Therefore the initial linear equations were:

$$.6815\mathbf{a} + \mathbf{b} = 400 \qquad (6\text{-}2)$$
$$.021\mathbf{a} + \mathbf{b} = 375 \qquad (6\text{-}3)$$

Solving for **a** and **b**, the results are **a** = 37.85 and **b** = 374.21. These values were then used with the Mathematics θ scores to transform all θ scores to Mathematics scaled scores.

$$\text{Mathematics Scaled Score} = 37.85(\theta) + 374.21 \qquad (6\text{-}4)$$

For Reading, the point on the θ scale where the standard setting committee decided that students had "met standard" was 1.127 and the point on the θ scale where the standard setting committee decided that students were "approaching standard" was -.227. Therefore the initial linear equations were:

$$1.127\mathbf{a} + \mathbf{b} = 400 \quad (6\text{-}5)$$
$$-.227\mathbf{a} + \mathbf{b} = 375 \quad (6\text{-}6)$$

Solving for a and b, the results are **a** = 18.46 and **b** = 379.19. These values were then used with the Reading θ scores to transform all θ scores to Reading scaled scores.

$$\text{Reading Scaled Score} = 18.46(\theta) + 379.19 \quad (6\text{-}7)$$

In Reading and Mathematics, students who earn scale scores below 375 are placed in "below standard, level 1" category in both Mathematics and Reading. Students who earn scale scores of 375 to 399 are placed in the "below standard, level 2" category in both Reading and Mathematics. Students who earn scale scores of 400 to 420 in Reading or 400 to 421 in Mathematics are in the "meets standard" category. Students who earn scale scores of 421 and higher in Reading or 422 and higher in Mathematics are in the "above standard" category.

### Listening and Writing

In the standard setting for Listening and Writing only a single cut score was set representing the standard. Therefore the linear transformations of θ for Listening and Writing required that one additional point be set. The decision was made to set the standard deviations of the θ scale of each test to a value so that the range of scale scores was within the 150 to 600 range obtained for the Reading and Mathematics tests. Once the linear transformation formula was obtained, all θ for the Listening and Writing tests were converted to whole number scaled scores. This now means that scale scores of 400 or higher meet the standard in all content areas and scale scores of 399 or lower are below the standard.

## CUT POINTS FOR CONTENT STRANDS

The cut points for the individual *content strands* in Reading and Mathematics were determined in the following manner. Using the θ value associated with "meets standard" and the item difficulties, it was possible to estimate the score of a proficient examinee on each of the items within the strand. Figure 6-1 gives a hypothetical distribution of item difficulties for the items in the Mathematics strands. As can be seen, the range of item difficulties differs for each strand. What may be less apparent is that the number of items below and above the theta value of .6815 also differs. Students receiving raw scores for each of the strands equal to or higher than the estimated strand score for proficient examinees are reported as "similar to the performance expected of students who met the standard". Raw score sums below this cut point are reported as "below the performance expected of students who met the standard". In Listening there are no scores reported at the strand level.

The Writing test consists of only two writing prompts, so using the partial credit model is not appropriate. Instead all scaling was done on the raw score scale. In Writing the cut-score for the two strands were determined in the following manner. The data from the standard-setting was divided into two sets, one consisting of examinees meeting the standard,

**Figure 6-1: Hypothetical Range of Item Difficulties (theta values) within Mathematics Strands**

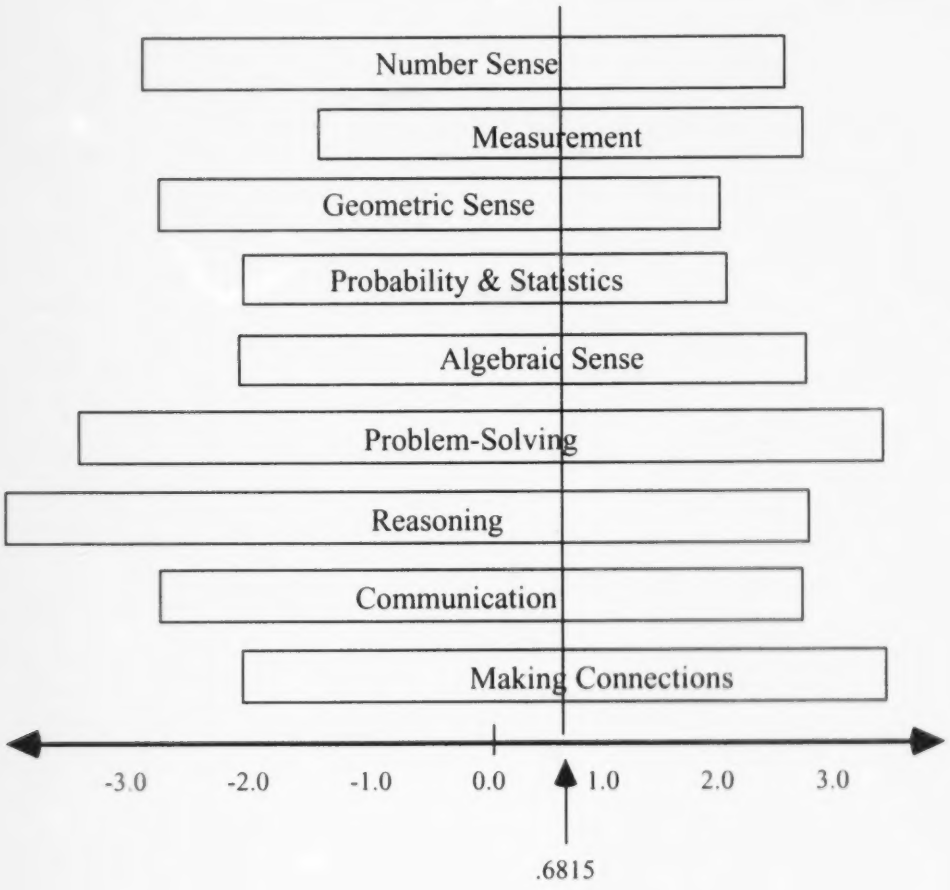
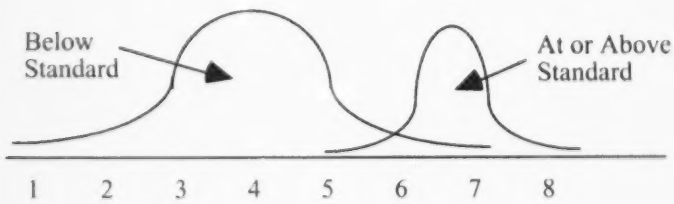



**Figure 6-2: Score Distribution of Students Identified as Below Standard and Score Distribution of Students Identified to Be At or Above Standard: Content, Organization, and Style**



the other examinees not meeting the standard. The raw scores for Writing Content, Organization, and Style and for Writing Mechanics were obtained for the examinees in each group (those meeting the standard and those not meeting the standard). Frequency distributions were computed on each of the strands for each group. Cut-points were identified as those showing the smallest overlap between the distributions of the two groups (see Figure 6-2). This is often referred to as a "contrasting groups design". Discussions of the standard setting committees also contributed to the decision. In the end, a minimum combined score of six for the Writing Content, Organization, and Style strand and a combined score of three for the Writing Mechanics strand were determined to be the cut points and the item parameters.

## EQUATING

The score scales established for the Grade 4 WASL in 1997 will stay in place for all subsequent years and test forms. Although new test forms are developed each year, Listening, Reading, and Mathematics are equated using items that were used in the base operational year (1997), thus maintaining the same scale score system, i.e., 400 for meeting the standard. Although the raw score to scale score relationship will change for Listening, Reading, and Mathematics, the level of difficulty associated with meeting the standard in each tested content area will remain statistically equivalent over time. The following is a summary of the procedures that are used for the equating of the Listening, Reading, and Mathematics tests of the Grade 4 WASL. The Writing test is not equated so the consistency across years is addressed during the training and scoring of the student papers. The same equating procedures and design will be used for Grades 7 and 10.

### Equating Reading and Mathematics Tests

In the description that follows, the process was completed separately for Reading and Mathematics; however, because the Reading and Mathematics tests are equated using the same design and procedure, the following description applies to both tests. In the first year of the operational assessment (1997), the multiple-choice, short-answer, and extended-response items were scaled using the Master's (1982) Partial Credit Model (PCM - see Pages 6-1 through 6-4 for a description of the model and the scaling process).

In order to equate the 1997 and 1998 test forms, anchor items were included in the Reading and Mathematics tests of each form. These items were common from one form to the next. The first step in performing the equating procedure was to evaluate the stability of anchor items over time. All items for a test (e.g., Mathematics) in a given form were calibrated to a PCM scale. Item difficulty estimates for the anchor items within each test were obtained for the 1997 form (from item calibrations in the summer of 1997) and for the 1998 form. The mean of the item difficulties for the anchor items was computed separately for each test form. The difference between the means was computed to establish an "equating constant." The equating constant was added to the item difficulties of each of the anchor items from the 1998 scaling, thus resulting in equal means for the anchor items on the two test forms.

Next, the item difficulty for each anchor item from the 1997 scaling was subtracted from the adjusted item difficulty for the same anchor item from the 1998 scaling. Any item

with an absolute difference greater than .3 was dropped from use as an anchor item (although these items were not dropped from the test and from the generation of test scores, score reports, etc.–they were simply no longer to be used as anchor items). If any items were dropped as anchor items, the computation of item difficulty means and equating constant, adjustment of item difficulties, and computation of differences in obtained and adjusted item difficulties was repeated. This process was repeated until the there was no loss of items.

Once a stable set of anchor items was obtained, the actual equating took place. This was done by analyzing the 1998 items for a test again, using the PCM, and fixing the item difficulties and step values for the valid anchor items to the values obtained on the 1997 test form. By fixing the item difficulties and step values of the anchor items, the resulting $\theta$ scale was the same in the 1998 test as it was for the 1997 test form. To derive the raw score to scale score relationship, the linear transformation equations for each test described on Pages 6-1 through 6-4 were used. This resulted in a consistent scale for each test across years.

**Equating the Listening Test**

Unlike the Reading and Mathematics tests, the Listening test is very short and consists of a single passage, read by the teacher, followed by six to eight items. This test design does not allow for the use of common items for equating. As a result, the contractor decided to use the anchor items from the Reading test for equating. This made the equating of the Listening test more complicated, involving more steps than needed for the Reading and Mathematics tests. A key component of this analysis was to make sure that the integrity of the Listening scale was maintained despite the use of the Reading common items.

Step 1. To begin with, the item difficulties for the Listening items were obtained from the 1997 testing. Holding these item difficulties and step values fixed, the Listening test items were rescaled including the Reading anchor items. This placed the Reading items on the Listening scale. This step was repeated for the 1998 test form, placing the Reading anchor items on the 1998 Listening scale.

Step 2. Using the Reading anchor item difficulties obtained in Step 1 for each form, it was possible to examine the stability of the common (anchor) items across forms. The same procedure outlined above for evaluating Reading and Mathematics anchor items was used to evaluate the Reading anchor items when projected onto the Listening scale.

Step 3. Once a set of stable anchor items was obtained for the 1997 to 1998 equating, the 1998 Listening test items were analyzed using the PCM and holding the item difficulties and step values for the anchor items fixed to those found for the 1997 test form described in Step 1. This produced item difficulties and step values for the current Listening test items that are on the same scale as the 1997 Listening scale.

Step 4. Using the item difficulties and step values obtained in Step 3, the raw score to $\theta$ scale values were obtained for the 1998 Listening test.

Step 5. The final raw score to scale score relationship for the 1998 Listening test was obtained by applying the same linear transformation used to obtain the raw score to scale score relationship for the 1997 form.

### Equating the Writing Test

For Writing, writing prompts were selected for the 1998 WASL that were of similar difficulty, purpose and audience as those from the 1997 WASL (difficulty assessed based on tryout data). The same scoring criteria were used to ensure constancy in writing difficulty. The raw score to scale score relationship did not change for the Writing test.

## NUMBER CORRECT SCORES TO SCALE SCORES

Each year WASL tests will have a different number correct score (raw score) to scale score relationship, although the underlying scale remains the same from year to year. This is possible because all items in the pool are on the same underlying Rasch scale. Table 6-1 gives the number correct score (NCS) to scale score (SS) relationship for the Listening, Reading, and Writing tests in the 1998 Grade 4 WASL. Table 6-2 gives the NCS to SS relationship for the 1998 Grade 4 Mathematics test.

**Table 6-1: 1998 Grade 4 Listening, Reading, and Writing Number Correct Scores (NCS) to Scale Scores (SS)**

| NCS | Listening SS | Reading SS | Writing SS |
|---|---|---|---|
| 0 | 192 | 300 | 213 |
| 1 | 224 | 313 | 233 |
| 2 | 262 | 327 | 254 |
| 3 | 291 | 336 | 275 |
| 4 | 316 | 342 | 296 |
| 5 | 342 | 347 | 317 |
| 6 | 369 | 351 | 338 |
| 7 | 400 | 355 | 358 |
| 8 | 433 | 359 | 379 |
| 9 | 482 | 362 | 400 |
| 10 | 520 | 365 | 421 |
| 11 | | 367 | 442 |
| 12 | | 370 | 463 |
| 13 | | 372 | |
| 14 | | 375 | |
| 15 | | 377 | |
| 16 | | 379 | |
| 17 | | 381 | |
| 18 | | 383 | |
| 19 | | 385 | |
| 20 | | 387 | |
| 21 | | 389 | |
| 22 | | 391 | |
| 23 | | 393 | |
| 24 | | 395 | |
| 25 | | 397 | |
| 26 | | 399 | |
| 27 | | 401 | |
| 28 | | 403 | |
| 29 | | 405 | |
| 30 | | 407 | |
| 31 | | 409 | |
| 32 | | 411 | |
| 33 | | 414 | |
| 34 | | 416 | |
| 35 | | 419 | |
| 36 | | 422 | |
| 37 | | 425 | |
| 38 | | 429 | |
| 39 | | 433 | |
| 40 | | 439 | |
| 41 | | 447 | |
| 42 | | 460 | |
| 43 | | 473 | |

**Table 6-2: 1998 Grade 4 Mathematics Number Correct Scores (NCS) to Scale Scores (SS)**

| NCS | Mathematics SS |
|---|---|
| 0 | 195 |
| 1 | 222 |
| 2 | 249 |
| 3 | 265 |
| 4 | 277 |
| 5 | 286 |
| 6 | 294 |
| 7 | 301 |
| 8 | 306 |
| 9 | 312 |
| 10 | 317 |
| 11 | 321 |
| 12 | 325 |
| 13 | 330 |
| 14 | 333 |
| 15 | 337 |
| 16 | 341 |
| 17 | 344 |
| 18 | 347 |
| 19 | 351 |
| 20 | 354 |
| 21 | 357 |
| 22 | 360 |
| 23 | 362 |
| 24 | 365 |
| 25 | 368 |
| 26 | 371 |
| 27 | 373 |
| 28 | 376 |
| 29 | 379 |
| 30 | 381 |
| 31 | 384 |

| NCS | Mathematics SS |
|---|---|
| 32 | 386 |
| 33 | 388 |
| 34 | 391 |
| 35 | 393 |
| 36 | 395 |
| 37 | 398 |
| 38 | 400 |
| 39 | 402 |
| 40 | 405 |
| 41 | 407 |
| 42 | 410 |
| 43 | 412 |
| 44 | 414 |
| 45 | 417 |
| 46 | 420 |
| 47 | 422 |
| 48 | 425 |
| 49 | 428 |
| 50 | 431 |
| 51 | 435 |
| 52 | 439 |
| 53 | 443 |
| 54 | 447 |
| 55 | 452 |
| 56 | 458 |
| 57 | 465 |
| 58 | 473 |
| 59 | 484 |
| 60 | 500 |
| 61 | 526 |
| 62 | 552 |

## Reference

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica*, (*47*), 149-174.

# PART 7

# RELIABILITY

The reliability of test scores is a measure of the degree to which the scores on the test are a "true" measure of the examinees' knowledge and skill relevant to the tested knowledge and skills. Simply put, the reliability is the proportion of observed score variance that is true score variance.

There are several ways to obtain estimates of score reliability: test-retest, alternate forms, internal consistency, and generalizability analysis are the most common. Test-retest estimates require administration of the same test at two different times. Typically the testing times for achievement tests are close together so that new learning does not impact scores. Alternate forms reliability estimates require administration of two parallel tests. These tests must be created in such a way that we have confidence that they measure the same domain of knowledge and skills using different items. Both test-retest and alternate forms estimates of score reliability require significant testing time for examinees and are generally avoided when there is a concern that fatigue or loss of motivation might impact the resulting reliability coefficient.

The *Washington Assessment of Student Learning* (WASL) is a rigorous measure that requires significant concentration on the part of students for a sustained period of time. For this reason, it was determined that test-retest and alternate forms reliability methods were unlikely to yield accurate estimates of score reliability. Therefore, an internal consistency measures were used to estimate score reliability for Reading, Listening, and Mathematics tests. A generalizability analysis was used to obtain an estimate of score reliability for the Writing test.

## INTERNAL CONSISTENCY AND GENERALIZABILITY

Internal consistency reliability is an indication of how similarly students perform across items measuring the same knowledge and skills—in other words, how consistent each examinee performs across all of the items within a test. Internal consistency can be estimated using Cronbach's alpha coefficient. When a test is composed entirely of multiple-choice (dichotomously scored) items, a modification of Cronbach's alpha can be used (KR-20). However, when multiple-point items are included on a test, Cronbach's alpha coefficient provides the internal consistency estimate. Two of the demands of applying this method when estimating score reliability are: 1) the number of items should be sufficient to obtain stable estimates of students' achievement and 2) all test items should be homogeneous (similar in type and measuring the very similar knowledge and skills).

WASL Reading and Mathematics tests have sufficient items to address the issue of test length; however, the Listening test has fewer items/scores, hence this will have a tendency to depress the alpha coefficient. WASL is also a complex measure that combines multiple-choice, short-answer, and extended response items. The Mathematics and Reading tests measure multiple strands that are all components of the domains of

Mathematics and Reading respectively. Hence, examinee performance may differ markedly from one item to another due to prior knowledge, educational experiences, exposure to similar content, etc. Because of this heterogeneity of items in the Reading and Mathematics tests and the short test length for the Listening test, use of Cronbach's alpha for estimating score reliability for WASL could result in an *under-estimate* of the reliability of scores. Generally it is believed that the true score reliability is higher than the estimate obtained through alpha when items are heterogeneous as they are in the WASL. The alpha coefficient is obtained through the following formula:

$$r_{xx'} = \left[\frac{N}{N-1}\right]\left(1 - \frac{\sum s_i^2}{s_x^2}\right)$$

Where:

$\sum s_i^2$ is the sum of all of the item variances

$\sum s_x^2$ is the observed score variance, and

N = the number of items on the test

For the Writing test, a generalizability analysis used to estimate reliability. This requires an analysis of all potential sources of score variance: raters, prompt, or individual differences, as well as interactions among these factors. Through an analysis of variance process, the total variance is partitioned into different sources. Variances due to rater, prompt, or interactions of rater by prompt, rater by examinee, and prompt by examinee are considered sources of error. In other words, the raters should not be systematically influenced by the written responses to a particular prompt, the raters should not be systematically affected by the examinees, examinees should not be systematically affected by an individual prompt, and raters, prompts, and examinees should not interact in any way. The estimate of reliability is the proportion of the observed score variance attributed to individual differences in performance rather than sources of error.

## STANDARD ERROR OF MEASUREMENT

One way to interpret the reliability of test scores is through the use of the Standard Error of Measurement ($s_{em}$). The $s_{em}$ is the standardized distribution of error around a given observed score. When one $s_{em}$ is added and subtracted from an observed score, we can be about 68 percent certain that the examinee's true score lies within the band. For example, if the $s_{em}$ for a test was 3.8, and the examinee's observed score was 32, we could be about 68 percent certain that the examinee's true score was between 32 - 3.8 and 32 + 3.8 or between 28.2 and 35.8. If we add and subtract two $s_{em}$, we can be about 95 percent certain that the examinee's true score lies within the band. Finally, if we add and subtract three $s_{em}$, we can be about 99 percent certain that the examinee's true score lies within the band. In classical testing, we obtain the $s_{em}$ through the following formula:

$$s_{em} = s_x \sqrt{1 - r_{xx'}}$$

Where:

$s_x$ is the observed score standard deviation, and

$r_{xx'}$ is the reliability estimate (alpha)

Table 7-1 provides the alpha coefficients and generalizability coefficient for each WASL test and the standard error of measurement for the scaled scores based on the standard deviation of the scale scores and the reliability coefficient.

**Table 7-1: 1998 Grade 4 Reliability Estimates and Standard Error Of Measurement for Each WASL Test**

| Subtest | Alpha Coefficient† or Generalizability Coefficient* | Scaled Score Standard Error of Measurement |
|---|---|---|
| Listening† | .60 | 35.84 |
| Reading† | .87 | 6.94 |
| Writing* | .70 | 24.09 |
| Mathematics† | .88 | 11.14 |

## INTERJUDGE AGREEMENT

As was described in Part 4, inter-judge (inter-rater) agreement consistency was another important source of evidence for the reliability of test scores. When two trained judges agree with the score given to a student's work, this gives support for the score on the short-answer or extended response item. Two methods are described in Part 4 for determining the degree to which judges gave equivalent score to the same student work: correlations between totals, when scores for open-ended items are summed, and percent agreement. Correlations between sums of open-ended item scores ranged from .96 to .99 across the Reading/Listening, Mathematics, and Writing tests. Exact agreement between two judges on scores for the Reading and Listening open-ended items ranged from 76 to 97 percent; exact and adjacent agreement ranged from 99 to 100 percent. Exact agreement between two judges on scores for the Mathematics open-ended items ranged from 79 to 96 percent; exact and adjacent agreement ranged from 95 to 99 percent. Exact agreement between two judges on scores for the Writing open-ended items ranged from 86 to 90 percent; exact and adjacent agreement was approximately 100 percent.

# PART 8

# DESCRIPTION OF PERFORMANCE FOR GRADE 4 STUDENTS

The data presented in this section of the report is descriptive of performance of Grade 4 students on the *Washington Assessment of Student Learning* (WASL) throughout the state. Included are means, standard deviations, and numbers tested for the all tested fourth graders and disaggregated by a variety of groups (Tables 8-1 through 8-14). Also presented are the percent of students in each gender, ethnic, and categorical program group who met or did not meet the standards for each content area (Tables 8-15 through 8-26). These data are useful for tracking, over time, the state's progress in helping students meet the Essential Academic Learning Requirements. One possible limitation to the data is that the categorization of students is based on the way students are classified on their response books. For example, when a school is identified as Title I-Schoolwide, all students should have been identified as Title I-Schoolwide on their answer documents. Finally, Tables 8-27 through 8-30 provide the mean performance on each item of the Grade 4 WASL tests, as well as the item-test correlations for each item.

## SUMMARY STATISTICS

The means for each score were computed by summing the relevant scores for all students tested and dividing by the total number of students tested. The standard deviation was computed by obtaining the square root of the relevant variances using the following equation:

$$SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

where:

$X$ is the individual score

$\bar{X}$ is the mean of scores for all students tested in the state, and

$N$ is the number of students tested in the state (those with valid scores)

Table 8-1 provides the state summary statistics for those Grade 4 students taking the WASL tests. The column headed "Points Possible" contains the maximum number of scale score points possible in each test for the 1998 form. The next two columns contain the mean scale score and standard deviation of the scale scores for all students tested in the state. Table 24 provides the state 1998 Grade 4 summary statistics for the WASL strands within tests. The column headed "Points Possible" indicates the maximum number of points possible in each strand for the 1998 form. The next two columns contain the mean number correct strand score and standard deviation of the strand scores for all students tested in the state. The final column indicates the percent of students whose performance on the strand was similar to those who met the standard. Tables 25 through 28 provide the summary data for each ethnic and gender group tested in 1998 (as indicated on the response book). Table 29 through 32 provide the summary data for students in each of the following categorical programs:

Learning Assistance Program (LAP) Reading, LAP Mathematics, Title I Reading, Title I Mathematics, Title I School, Bilingual/English as a Second Language (ESL), Highly Capable Students, Section 504, Special Education, and Migrant Education (as indicated on the response book).

**Table 8-1: 1998 Grade 4 Scale Score Means, Standard Deviations, and Maximum Scale Scores by Test**

| Test | Number Tested | Maximum Scale Score | Mean Scale Score | Standard Deviation |
|---|---|---|---|---|
| Listening | 72994 | 520 | 414.52 | 56.67 |
| Reading | 72473 | 473 | 402.12 | 19.24 |
| Writing | 68631 | 463 | 376.46 | 43.99 |
| Mathematics | 73164 | 552 | 383.50 | 32.16 |

**Table 8-2: 1998 Grade 4 Maximum Number Possible, Number Correct Score Means, Standard Deviations (SD) by Strand, and Percent of Students with Strength in Strand**

| Strand | Number Tested | Points Possible | Mean | SD | Percent with Strength in Strand |
|---|---|---|---|---|---|
| Main Ideas & Details of Fiction | 72473 | 8 | 5.68 | 1.79 | 58.6 |
| Analysis, Interpretation, Critique of Fiction | 72473 | 10 | 6.50 | 2.65 | 53.8 |
| Main Ideas & Details of Non-fiction Text | 72473 | 9 | 6.25 | 1.97 | 47.9 |
| Analysis, Interpretation, Critique of Non-fiction Text | 72473 | 16 | 8.57 | 3.28 | 49.6 |
| Writing Content, Organization & Style | 68631 | 8 | 5.09 | 1.39 | 37.5 |
| Writing Mechanics | 68631 | 4 | 2.77 | 1.00 | 49.0 |
| Number Sense | 73164 | 8 | 4.06 | 1.66 | 35.4 |
| Measurement | 73164 | 7 | 4.74 | 1.55 | 33.8 |
| Geometric Sense | 73164 | 7 | 3.92 | 1.69 | 36.2 |
| Probability & Statistics | 73164 | 6 | 3.28 | 1.46 | 45.0 |
| Algebraic Sense | 73164 | 6 | 3.10 | 1.61 | 39.2 |
| Solves Problems | 73164 | 5 | 1.95 | 1.46 | 32.3 |
| Reasons Logically | 73164 | 12 | 5.35 | 3.31 | 36.0 |
| Communicates Understanding | 73164 | 7 | 3.18 | 1.85 | 36.3 |
| Makes Connections | 73164 | 4 | 1.95 | 1.10 | 30.6 |

**Table 8-3: 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender**

| Gender | Number Tested | Mean | SD |
|---|---|---|---|
| Males | 37369 | 415.39 | 55.73 |
| Females | 35562 | 413.64 | 57.61 |

**Table 8-4: 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group**

| Ethnic Group | Number Tested | Mean | SD |
|---|---|---|---|
| African American/Black | 3498 | 397.49 | 58.14 |
| Alaska Native/Native American | 1966 | 398.27 | 58.49 |
| Asian/Pacific Islander | 5097 | 408.71 | 58.69 |
| Latino/Hispanic | 6207 | 382.64 | 62.32 |
| White/Caucasian | 54769 | 420.36 | 53.98 |
| Multi-Racial | 831 | 415.30 | 57.09 |

**Table 8-5: 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender**

| Gender | Number Tested | Mean | SD |
|---|---|---|---|
| Males | 37002 | 400.02 | 19.27 |
| Females | 35409 | 404.32 | 18.95 |

**Table 8-6: 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group**

| Ethnic Group | Number Tested | Mean | SD |
|---|---|---|---|
| African American/Black | 3457 | 393.37 | 17.95 |
| Alaska Native/Native American | 1936 | 392.29 | 18.23 |
| Asian/Pacific Islander | 5071 | 401.79 | 19.00 |
| Latino/Hispanic | 6128 | 389.14 | 18.80 |
| White/Caucasian | 54437 | 404.57 | 18.56 |
| Multi-Racial | 824 | 399.63 | 18.29 |

**Table 8-7: 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender**

| Gender | Number Tested | Mean | SD |
|---|---|---|---|
| Males | 34987 | 368.37 | 43.42 |
| Females | 33594 | 384.91 | 42.98 |

**Table 8-8: 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group**

| Gender or Ethnic Group | Number Tested | Mean | SD |
|---|---|---|---|
| African American/Black | 3214 | 364.68 | 42.89 |
| Alaska Native/Native American | 1808 | 356.85 | 44.77 |
| Asian/Pacific Islander | 4812 | 384.31 | 42.63 |
| Latino/Hispanic | 5577 | 352.70 | 44.88 |
| White/Caucasian | 51858 | 379.92 | 42.86 |
| Multi-Racial | 776 | 367.95 | 44.28 |

**Table 8-9: 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Gender**

| Gender | Number Tested | Mean | SD |
|---|---|---|---|
| Males | 37448 | 383.67 | 33.04 |
| Females | 35654 | 383.34 | 31.21 |

**Table 8-10: 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Ethnic Group**

| Ethnic Group | Number Tested | Mean | SD |
|---|---|---|---|
| African American/Black | 3512 | 366.19 | 30.03 |
| Alaska Native/Native American | 1972 | 366.77 | 29.81 |
| Asian/Pacific Islander | 5104 | 385.09 | 33.63 |
| Latino/Hispanic | 6240 | 361.90 | 30.60 |
| White/Caucasian | 54877 | 387.60 | 30.87 |
| Multi-Racial | 832 | 378.74 | 29.51 |

**Table 8-11: 1998 Grade 4 Listening Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program**

| Categorical Program | Number Tested | Mean | SD |
|---|---|---|---|
| LAP Reading | 2477 | 389.15 | 57.23 |
| LAP Mathematics | 2425 | 387.51 | 55.51 |
| Title I Reading | 3988 | 388.68 | 56.97 |
| Title I Mathematics | 1940 | 387.47 | 57.37 |
| Title I School-wide | 9632 | 393.96 | 60.13 |
| Section 504 | 386 | 398.35 | 56.88 |
| Special Education | 6840 | 377.40 | 61.75 |
| Title I Migrant Education | 564 | 361.09 | 60.67 |
| Bilingual/ESL | 3120 | 361.94 | 59.05 |
| Gifted/Highly Capable Students | 3214 | 451.80 | 45.57 |

**Table 8-12: 1998 Grade 4 Reading Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program**

| Categorical Program | Number Tested | Mean | SD |
|---|---|---|---|
| LAP Reading | 2465 | 387.41 | 14.78 |
| LAP Mathematics | 2409 | 387.57 | 15.54 |
| Title I Reading | 3947 | 386.56 | 14.60 |
| Title I Mathematics | 1883 | 386.96 | 14.90 |
| Title I School-wide | 9510 | 393.13 | 18.85 |
| Section 504 | 382 | 393.46 | 17.86 |
| Special Education | 6513 | 381.00 | 17.54 |
| Title I Migrant Education | 555 | 380.84 | 16.44 |
| Bilingual/ESL | 3061 | 381.86 | 16.08 |
| Gifted/Highly Capable Students | 3215 | 424.07 | 14.99 |

**Table 8-13: 1998 Grade 4 Writing Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical Program**

| Categorical Program | Number Tested | Mean | SD |
|---|---|---|---|
| LAP Reading | 2272 | 348.99 | 37.74 |
| LAP Mathematics | 2218 | 346.89 | 37.99 |
| Title I Reading | 3652 | 344.93 | 36.59 |
| Title I Mathematics | 1735 | 345.74 | 36.62 |
| Title I School-wide | 8837 | 360.36 | 44.29 |
| Section 504 | 359 | 353.35 | 40.72 |
| Special Education | 5922 | 334.98 | 41.86 |
| Title I Migrant Education | 482 | 333.67 | 42.26 |
| Bilingual/ESL | 2685 | 342.00 | 41.43 |
| Gifted/Highly Capable Students | 3125 | 418.50 | 34.06 |

**Table 8-14: 1998 Grade 4 Mathematics Test: Number Tested, Scale Score Means, and Standard Deviations (SD) by Categorical**

| Categorical Program | Number Tested | Mean | SD |
|---|---|---|---|
| LAP Reading | 2494 | 360.88 | 27.29 |
| LAP Mathematics | 2436 | 357.72 | 25.48 |
| Title I Reading | 3992 | 359.80 | 25.91 |
| Title I Mathematics | 1908 | 357.83 | 24.72 |
| Title I School-wide | 9713 | 368.46 | 31.38 |
| Section 504 | 388 | 371.36 | 33.02 |
| Special Education | 6804 | 355.22 | 31.25 |
| Title I Migrant Education | 568 | 350.86 | 28.11 |
| Bilingual/ESL | 3149 | 354.25 | 28.99 |
| Gifted/Highly Capable Students | 3221 | 423.55 | 26.83 |

## PERCENT MEETING STANDARD

Tables 8-15 through 8-22 provide the information regarding the number of students in the state as well as in each gender and ethnic group who met the standard in Listening, Reading, Writing, and Mathematics. Tables 23 through 30 provide the information regarding the number of students in each categorical program who met the standard in Listening, Reading, Writing, and Mathematics. These data can be monitored to determine whether the percent achieving the standards improves over time. The following are the general descriptions of the performance levels established for the Washington Assessment of Student Learning:

Level 4 Above Standard: This level represents superior performance, notably above that required for meeting the standard at grade 4.

Level 3 MEETS STANDARD*: This level represents solid academic performance for grade 4. Students reaching this level have demonstrated proficiency over challenging content, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate for the content and grade level.

Level 2 Below Standard: This level denotes partial accomplishment of the knowledge and skills that are fundamental for meeting the standard at grade 4.

Level 1 Well Below Standard: This level denotes little or no demonstration of the prerequisite knowledge and skills that are fundamental for meeting the standard at grade 4.

* *In all content areas, "Meets Standard" reflects what a well taught, hard working student should know and be able to do.*

For the Writing and Listening tests, the tables show, for each group, the percent meeting standard, the percent not meeting standard, and the percent of students exempted. For the Reading and Mathematics tests, the tables show, for each group, the percent in each performance level and the percent exempted. For Reading and Mathematics, students in Levels 1 and 2 did not meet the standard. Students in Levels 3 and 4 met or exceeded the standard.

**Table 8-15: 1998 Grade 4 Listening Test: Percent Meeting Standards by Total and by Gender**

| Group | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| All Students | 73,377 | 71.3 | 27.0 | 1.7 |
| Males | 37,488 | 70.5 | 27.6 | 1.9 |
| Females | 35,655 | 69.8 | 28.7 | 1.5 |

**Table 8-16: 1998 Grade 4 Listening Test: Percent Meeting Standards by Ethnic Group**

| Ethnic Group | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| African American/Black | 3,507 | 58.2 | 38.9 | 2.8 |
| Alaska Native/Native American | 1,970 | 57.8 | 39.7 | 2.4 |
| Asian/Pacific Islander | 5,110 | 66.3 | 31.9 | 1.8 |
| Latino/Hispanic | 6,23[illegible] | 46.6 | 50.1 | 3.3 |
| White/Caucasian | 54,859 | 74.5 | 24.1 | 1.4 |
| Multi-Racial | 837 | 70.2 | 26.9 | 0.2 |

**Table 8-17: 1998 Grade 4 Reading Test: Percent Meeting Standards by Total and by Gender**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Group | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| All Students | 72,853 | 15.6 | 40.0 | 34.6 | 7.5 | 1.9 |
| Males | 36,684 | 12.8 | 37.3 | 36.6 | 11.2 | 2.1 |
| Females | 35,402 | 18.0 | 40.0 | 34.6 | 7.5 | 1.6 |

**Table 8-18: 1998 Grade 4 Reading Test: Percent Meeting Standards by Ethnic Group (1998)**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Ethnic Group | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| African American/Black | 3,449 | 5.9 | 28.2 | 46.4 | 16.4 | 3.2 |
| Alaska Native/Native American | 1,929 | 5.2 | 26.7 | 47.3 | 18.1 | 2.8 |
| Asian/Pacific Islander | 5,076 | 15.0 | 37.9 | 37.2 | 8.0 | 1.9 |
| Latino/Hispanic | 6,112 | 4.5 | 22.1 | 46.8 | 23.0 | 3.6 |
| White/Caucasian | 54,375 | 17.7 | 42.7 | 31.0 | 7.0 | 1.6 |
| Multi-Racial | 828 | 11.3 | 38.6 | 38.1 | 11.6 | 0.4 |

**Table 8-19: 1998 Grade 4 Writing Test: Percent Meeting Standards by Total and Gender**

| Group | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| All Students | 68,981 | 36.7 | 61.3 | 2.0 |
| Males | 34,504 | 28.8 | 68.8 | 2.3 |
| Females | 33,262 | 43.6 | 54.7 | 1.7 |

**Table 8-20: 1998 Grade 4 Writing Test: Percent Meeting Standards by Ethnic Group**

| Ethnic Group | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| African American/Black | 3,171 | 24.5 | 72.0 | 3.5 |
| Alaska Native/Native American | 1,761 | 20.7 | 76.2 | 3.1 |
| Asian/Pacific Islander | 4,777 | 42.8 | 55.1 | 2.0 |
| Latino/Hispanic | 5,441 | 17.7 | 78.4 | 3.9 |
| White/Caucasian | 51,271 | 39.0 | 59.3 | 1.7 |
| Multi-Racial | 765 | 29.7 | 69.8 | 0.5 |

**Table 8-21: 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Total and Gender**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Group | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| All Students | 73,566 | 10.8 | 19.8 | 29.2 | 38.2 | 2.0 |
| Males | 36,753 | 11.6 | 19.6 | 28.0 | 38.6 | 2.3 |
| Females | 35,225 | 10.0 | 20.1 | 30.5 | 37.8 | 1.7 |

**Table 8-22: 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Ethnic Group**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Group | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| African American/Black | 3,428 | 2.9 | 9.6 | 24.0 | 60.1 | 3.4 |
| Alaska Native/Native American | 1,905 | 3.7 | 9.6 | 24.1 | 59.6 | 3.1 |
| Asian/Pacific Islander | 5,053 | 12.7 | 20.0 | 27.8 | 37.4 | 2.0 |
| Latino/Hispanic | 6,048 | 2.7 | 8.2 | 20.6 | 64.6 | 3.9 |
| White/Caucasian | 54,113 | 12.4 | 22.3 | 30.8 | 32.8 | 1.7 |
| Multi-Racial | 816 | 6.6 | 18.0 | 30.6 | 44.1 | 0.7 |

**Table 8-23: 1998 Grade 4 Listening Test: Percent Meeting Standards by Categorical Program**

| Categorical Program | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| LAP Reading | 1,250 | 54.0 | 45.6 | 0.4 |
| LAP Mathematics | 1,938 | 53.0 | 46.6 | 0.4 |
| Title I Reading | 4,004 | 53.6 | 45.9 | |
| Title I Mathematics | 1,910 | 51.9 | 47.4 | 0.7 |
| Title I School-wide | 8,656 | 56.4 | 41.9 | 1.7 |
| Section 504 | 389 | 59.4 | 37.3 | 3.2 |
| Special Education | 6,872 | 43.3 | 51.2 | 5.5 |
| Title I Migrant Education | 571 | 33.1 | 63.0 | 3.9 |
| Bilingual/ESL | 3,132 | 32.4 | 62.3 | 5.4 |
| Gifted/Highly Capable Students | 3,220 | 92.8 | 7.1 | 0.1 |

**Table 8-24: Grade 4 Reading Test: Percent Meeting Standards by Categorical Program**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Categorical Program | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| LAP Reading | 1,233 | 1.5 | 17.8 | 62.2 | 18.0 | 0.5 |
| LAP Mathematics | 1,917 | 1.7 | 20.1 | 58.4 | 19.4 | 0.6 |
| Title I Reading | 3,953 | 1.3 | 15.5 | 63.7 | 18.8 | 0.7 |
| Title I Mathematics | 1,883 | 1.1 | 17.6 | 61.3 | 19.0 | 1.0 |
| Title I School-wide | 9,509 | 6.5 | 28.1 | 45.5 | 17.9 | 2.0 |
| Section 504 | 385 | 5.7 | 28.2 | 48.8 | 14.1 | 3.2 |
| Special Education | 6,493 | 1.4 | 11.4 | 43.9 | 36.5 | 6.9 |
| Title I Migrant Education | 559 | 1.0 | 9.7 | 50.5 | 33.9 | 4.9 |
| Bilingual/ESL | 3,050 | 1.1 | 10.8 | 51.2 | 31.1 | 5.9 |
| Gifted/Highly Capable | 3,214 | 59.0 | 36.8 | 3.7 | 0.4 | 0.1 |

**Table 8-25: 1998 Grade 4 Writing Test: Percent Meeting Standards by Categorical Program**

| Categorical Program | Number of Students | Percent Meeting Standard | Percent Not Meeting Standard | Percent Exempt |
|---|---|---|---|---|
| LAP Reading | 1,102 | 12.5 | 86.9 | 0.6 |
| LAP Mathematics | 1,742 | 11.7 | 87.6 | 0.7 |
| Title I Reading | 3,593 | 9.8 | 89.4 | 0.8 |
| Title I Mathematics | 1,704 | 10.6 | 88.3 | 1.1 |
| Title I School-wide | 8,656 | 22.9 | 75.1 | 2.0 |
| Section 504 | 358 | 18.8 | 77.9 | 3.2 |
| Special Education | 5,680 | 7.3 | 86.0 | 6.7 |
| Title I Migrant Education | 608 | 6.4 | 89.0 | 4.6 |
| Bilingual/ESL | 2,603 | 10.0 | 84.1 | 5.9 |
| Gifted/Highly Capable Students | 3,109 | 78.8 | 21.1 | 0.1 |

**Table 8-26: 1998 Grade 4 Mathematics Test: Percent Meeting Standards by Categorical Program**

| | | Meets Standard | | Does Not Meet Standard | | |
|---|---|---|---|---|---|---|
| Categorical Program | No. of Students | Percent Level 4 | Percent Level 3 | Percent Level 2 | Percent Level 1 | Percent Exempt |
| LAP Reading | 1,224 | 1.5 | 6.3 | 21.8 | 69.9 | 0.6 |
| LAP Mathematics | 1,906 | 0.4 | 4.3 | 20.2 | 74.5 | 0.6 |
| Title I Reading | 3,918 | 1.1 | 5.0 | 21.1 | 72.2 | 0.6 |
| Title I Mathematics | 1,866 | 0.5 | 3.6 | 20.2 | 74.9 | 0.8 |
| Title I School-wide | 9,416 | 4.5 | 11.7 | 23.7 | 58.1 | 2.0 |
| Section 504 | 385 | 6.4 | 14.1 | 22.3 | 54.2 | 3.0 |
| Special Education | 6,412 | 1.5 | 5.8 | 16.4 | 69.4 | 6.8 |
| Title I Migrant Education | 548 | 1.0 | 2.8 | 14.6 | 77.1 | 4.4 |
| Bilingual/ESL | 3,014 | 1.5 | 4.8 | 14.5 | 73.5 | 5.6 |
| Gifted/Highly Capable | 3,207 | 51.7 | 32.9 | 12.2 | 3.1 | 0.1 |

## MEAN ITEM PERFORMANCE AND ITEM-TEST CORRELATIONS

As discussed in Part 2, traditional item statistics were used, along with Rasch difficulties and fit statistics, to evaluate the quality of items. All items in the pool were evaluated together and items that met quality standards were retained in the item pool. Mean item performance for multiple choice items can range from 0 to 1. This is often called the p-value. Mean item performance for short-answer items can range from 0 to 2. Mean item performance for extended response items can range from 0 to 4. For the Writing test, mean scores represent the average scores for each of the scoring rules applied to the written piece. There are two written pieces in the Grade 4 WASL. Students can receive from 0 to 4 points for Content, Organization, and Style and from 0 to 2 points for Writing Mechanics for *each* of the written pieces. The higher the mean item performance, the easier the item. Item-test correlations can range from -1.0 to 1.0; positive correlations indicate that item performance is related to overall test performance. Rasch item difficulties can range from –4.0 to 4.0, with negative numbers representing easier items and positive numbers representing more difficult items. The data provided in Tables 8-27 through 8-30 indicate the number of points possible for the items or writing scores, the item or score means, the item score to test score correlations, and the Rasch item difficulties for each of the items in the Listening, Writing, Reading, and Mathematics tests respectively.

**Table 8-27: 1998 Grade 4 Listening Test: Number of Points Possible Per Item, Mean Item Performance, and Item-Test Correlation for Each Item**

| Item Number in Test Booklet | Number Possible | Item Mean | Item-Test Correlation | Rasch Item Difficulty |
|---|---|---|---|---|
| 1 | 2 | 1.03 | 0.20 | 1.19 |
| 2 | 1 | 0.68 | 0.28 | 0.41 |
| 3 | 1 | 0.91 | 0.35 | -1.40 |
| 4 | 1 | 0.91 | 0.32 | -1.39 |
| 5 | 1 | 0.90 | 0.31 | -1.24 |
| 6 | 2 | 0.81 | 0.13 | 1.80 |
| 7 | 1 | 0.96 | 0.18 | -2.29 |
| 8 | 1 | 0.90 | 0.37 | -1.22 |

**Table 8-28: 1998 Grade 4 Writing Test: Number of Points Possible Per Score-Type, Mean Score, and Score-Total Test Correlation for Each Score**

| Prompt Number | Score Type | Score Points Possible | Score Mean | Item-Test Correlation | Rasch Item Difficulty |
|---|---|---|---|---|---|
| 1 | Content, Organization & Style | 4 | 2.58 | 0.57 | -0.94 |
| | Writing Mechanics | 2 | 1.31 | 0.62 | 0.08 |
| 2 | Content, Organization & Style | 4 | 2.33 | 0.49 | 0.36 |
| | Writing Mechanics | 2 | 1.24 | 0.59 | 0.50 |

**Table 8-29: 1998 Grade 4 Reading Test: Number of Points Possible Per Item, Mean Item Performance, Item-Test Correlation, and Rasch Item Difficulty for Each Item**

| Item Number in Test Booklet | Points Possible | Item Mean | Item-Test Correlation | Rasch Item Difficulty |
|---|---|---|---|---|
| 1 | 1 | 0.65 | 0.28 | 0.47 |
| 2 | 1 | 0.61 | 0.44 | 0.67 |
| 3 | 1 | 0.86 | 0.30 | -0.85 |
| 4 | 1 | 0.55 | 0.23 | 0.97 |
| 5 | 2 | 1.06 | 0.37 | 1.11 |
| 6 | 2 | 1.46 | 0.37 | -0.06 |
| 7 | 1 | 0.90 | 0.34 | -1.39 |
| 8 | 1 | 0.73 | 0.37 | 0.02 |
| 9 | 1 | 0.93 | .041 | -1.56 |
| 10 | 1 | 0.71 | 0.56 | 0.20 |
| 11 | 1 | 0.65 | 0.48 | 0.50 |
| 12 | 1 | 0.76 | 0.42 | -0.14 |
| 13 | 1 | 0.77 | 0.44 | -0.22 |
| 14 | 2 | 1.33 | 0.57 | 0.39 |
| 15 | 1 | 0.52 | 0.28 | 1.12 |
| 16 | 2 | 1.30 | 0.55 | 0.66 |
| 17 | 4 | 2.17 | 0.64 | 1.09 |
| 18 | 1 | 0.90 | 0.32 | -1.34 |
| 19 | 2 | 1.04 | 0.59 | 1.16 |
| 20 | 1 | 0.73 | 0.45 | 0.07 |
| 21 | 1 | 0.76 | 0.36 | -0.14 |
| 22 | 2 | 0.70 | 0.36 | 1.82 |
| 23 | 4 | 2.23 | 0.51 | 0.97 |
| 24 | 1 | 0.71 | 0.33 | 0.17 |
| 25 | 2 | 1.14 | 0.54 | 0.83 |
| 26 | 1 | 0.64 | 0.43 | 0.56 |
| 27 | 1 | 0.81 | 0.39 | -0.46 |
| 28 | 1 | 0.77 | 0.42 | -0.13 |
| 29 | 1 | 0.42 | 0.20 | 1.64 |
| 30 | 1 | 0.41 | 0.23 | 1.59 |

**Table 8-30: 1998 Grade 4 Mathematics Test: Number of Points Possible Per Item, Mean Item Performance, Item-Test Correlation, and Rasch Item Difficulty for Each Item**

| Item Number in Test Booklet | Points Possible | Item Mean | Item-Test Correlation | Rasch Item Difficulty |
|---|---|---|---|---|
| 1 | 1 | 0.72 | 0.33 | -0.76 |
| 2 | 1 | 0.89 | 0.40 | -2.05 |
| 3 | 2 | 1.43 | 0.37 | -0.70 |
| 4 | 1 | 0.88 | 0.24 | -1.93 |
| 5 | 2 | 0.53 | 0.43 | 1.00 |
| 6 | 1 | 0.76 | 0.34 | -1.01 |
| 7 | 1 | 0.58 | 0.38 | 0.05 |
| 8 | 2 | 1.57 | 0.35 | -0.87 |
| 9 | 1 | 0.42 | 0.32 | 0.72 |
| 10 | 4 | 1.26 | 0.46 | -0.92 |
| 11 | 1 | 0.68 | 0.27 | -0.54 |
| 12 | 1 | 0.74 | 0.30 | -0.88 |
| 13 | 1 | 0.49 | 0.37 | 0.33 |
| 14 | 2 | 0.92 | 0.53 | 0.74 |
| 15 | 1 | 0.45 | 0.32 | 0.66 |
| 16 | 1 | 0.50 | 0.34 | 0.31 |
| 17 | 1 | 0.43 | 0.34 | 0.65 |
| 18 | 4 | 2.13 | 0.61 | 0.33 |
| 19 | 1 | 0.33 | 0.09 | 1.02 |
| 20 | 2 | 0.92 | 0.51 | 0.50 |
| 21 | 1 | 0.68 | 0.32 | -0.53 |
| 22 | 1 | 0.51 | 0.24 | 0.26 |
| 23 | 2 | 0.97 | 0.54 | 0.39 |
| 24 | 1 | 0.26 | 0.26 | 1.49 |
| 25 | 2 | 1.78 | 0.35 | -1.29 |
| 26 | 2 | 1.49 | 0.41 | -0.52 |
| 27 | 1 | 0.42 | 0.29 | 0.70 |
| 28 | 1 | 0.69 | 0.33 | -0.60 |
| 29 | 2 | 1.20 | 0.50 | -0.08 |
| 30 | 1 | 0.59 | 0.34 | 0.00 |
| 31 | 1 | 0.40 | 0.28 | 0.68 |
| 32 | 2 | 0.51 | 0.48 | 1.20 |
| 33 | 1 | 0.72 | 0.42 | -0.83 |
| 34 | 2 | 0.67 | 0.53 | 1.05 |
| 35 | 2 | 0.78 | 0.49 | 0.78 |
| 36 | 1 | 0.53 | 0.31 | 0.20 |
| 37 | 4 | 1.68 | 0.57 | 0.56 |
| 38 | 1 | 0.32 | 0.30 | 1.24 |
| 39 | 2 | 1.71 | 0.29 | -1.28 |
| 40 | 1 | 0.53 | -0.03 | 0.21 |

# APPENDIX A

## Washington Essential Academic Learning Requirements in Reading, Writing, Communication, and Mathematics

## Reading

1. **The student understands and uses different skills and strategies to read.**
   - 1.1 Uses word recognition and word meaning skills to read and comprehend text (e.g., phonics, context clues, picture clues, and word origins; roots, prefixes, and suffixes of words)
   - 1.2 Builds vocabulary through reading
   - 1.3 Reads fluently, adjusting reading for purpose and material
   - 1.4 Understands elements of literary (fiction)
   - 1.5 Understands features of non-fiction text and computer software (e.g., titles, headings, pictures, maps, and charts to find and understand specific information)

2. **The student understands the meaning of what is read.**
   - 2.1 Comprehends important ideas and details
   - 2.2 Expands comprehension by analyzing, synthesizing, and interpreting information and ideas
   - 2.3 Thinks critically about text and analyzes author's use of language, style, purpose, and perspective

3. **The student reads different materials for a variety of purposes.**
   - 3.1 Reads to learn new information
   - 3.2 Reads to perform tasks
   - 3.3 Reads for literary experience
   - 3.4 Reads for career applications

4. **The student sets goals and evaluates progress to improve reading.**
   - 4.1 Assesses strengths and need for improvement
   - 4.2 Seeks and offers feedback to improve reading
   - 4.3 Develops interests and shares reading experiences

## Writing

1. **The student writes clearly and effectively**
   - 1.1 Develops concept and design (develops a topic or theme; organizes written thoughts with a clear beginning, middle, and end; uses transitional sentences and phrases to connect ideas; writes coherently and effectively)
   - 1.2 Uses style appropriate to audience and purpose (uses voice, word choice, and sentence fluency for intended style and audience)
   - 1.3 Applies writing conventions (grammar, punctuation, capitalization)

2. **The student writes in a variety of forms for different audiences and purposes.**
   - 2.1 Writes for different audiences
   - 2.2 Writes for different purposes (telling stories, presenting analytical responses to literature, persuading, conveying technical information, completing a team project, explaining concepts and procedures)
   - 2.3 Writes in a variety of forms (narratives, journals, poems, essays, stories, research reports, and technical writing)

3. **The student understands and uses the steps of the writing process.**
   - 3.1 Prewrites (generates ideas and gather information for writing)
   - 3.2 Drafts (elaborates on a topic and supporting ideas)
   - 3.3 Revises (collects input and enhances style and text)
   - 3.4 Edits (uses resources to correct spelling, punctuation, grammar, and usage)
   - 3.5 Publishes (selects publishing form and produces a completed writing project to share with a chosen audience)

4. **The student analyzes and evaluates the effectiveness of written work.**
   - 4.1 Assesses own strengths and needs for improvement (analyzes effectiveness of own writing and sets goals for improvement)
   - 4.2 Seeks and offers feedback

## Communication

1. **The student uses listening and observing skills to gain understanding.**
   - 1.1 Focuses attention
   - 1.2 Listens and observes to gain and interpret information
   - 1.3 Checks for understanding by asking questions and paraphrasing

2. **The student communicates ideas clearly and effectively.**
   - 2.1 Communicates clearly to a range of audiences for different purposes
   - 2.2 Develops content and ideas (develops a topic or theme; organizes thoughts around a clear beginning, middle, and end; uses transitional sentences and phrases to connect ideas; speaks coherently and effectively)
   - 2.3 Uses effective delivery (adjusts speaking strategies for a variety of audiences and purposes by varying tone, pitch, projection, posture, eye contact, facial expressions body language, and pace of speech to create effect and aid communication)
   - 2.4 Uses effective language and style (uses language that is grammatically correct, precise, engaging, and well suited to topic, audience and purpose)
   - 2.5 Effectively uses action, sound, and/or images to support presentations

3. **The student uses communication strategies and skills to work effectively with others.**
   - 3.1 Uses language to interact effectively and responsibly with others
   - 3.2 Works cooperatively as a member of a group
   - 3.3 Seeks agreement and solutions through discussion

4. **The student analyzes and evaluates the effectiveness of formal and informal communication.**
   - 4.1 Assess strengths and needs for improvement (analyzes effectiveness of own writing and sets goals for improvement)
   - 4.2 Seeks and offers feedback (seeks and uses feedback to improve communication; offers suggestions and comments to others)
   - 4.3 Analyzes mass communication
   - 4.4 Analyzes how communication is used in career settings

## Mathematics

1. **The student understands and applies the concepts and procedures of mathematics.**
   - 1.1 Understands and applies concepts and procedures of number sense (number and numeration, number theory, computation, and estimation)
   - 1.2 Understands and applies concepts and procedures of measurement (attributes and dimensions, approximation and precision, systems and tools)
   - 1.3 Understands and applies concepts and procedures of geometric sense (shape and dimension, relationships, and transformation)
   - 1.4 Understands and applies concepts and procedures of probability and statistics (probability, statistics, prediction, and inference)
   - 1.5 Understands and applies concepts and procedures of algebraic sense (patterns, relations, representations, and operations)

2. **The student uses mathematics to define and solve problems.**
   - 2.1 Investigates situations (by searching for patterns and exploring a variety of approaches)
   - 2.2 Formulates questions and defines problem
   - 2.3 Constructs solutions (by choosing necessary information and using the appropriate tools, concepts and procedures)

3. **The student uses mathematical reasoning.**
   - 3.1 Analyzes information (from a variety of sources; uses models, known facts, patterns, and relationships to validate thinking)
   - 3.2 Predicts results and makes inferences and conjectures based on analysis of problem situations
   - 3.3 Draws conclusions and verifies results (supports mathematical arguments, justifies results, and checks for reasonableness of solutions)

4. **The student uses communicates knowledge and understanding in both everyday and mathematical language.**
   - 4.1 Gathers information (reads, listens, and observes to extract mathematical information)
   - 4.2 Organizes and interprets information
   - 4.3 Represents and shares mathematical information (shares, explains, defends mathematical ideas using terms, language, charts, and graphs, etc. that can be clearly understood by a variety of audiences

## Mathematics (Cont.)

4. **The student understands how mathematical ideas connect within mathematics, to other subject areas, and to real life situations.**
   - 5.1 Relates ideas and concepts within mathematics (recognizes relationships among mathematical ideas and topics)
   - 5.2 Relates mathematical concepts and procedures to other disciplines (identifies and applies mathematical thinking and notation in other subject areas
   - 5.3 Relates mathematical concepts and procedures to real-life situations (understands the connections between mathematics and problem solving skills used every day at work and at home)

# APPENDIX B

# WASHINGTON ASSESSMENT OF STUDENT LEARNING

# GRADE 4

# MATHEMATICS TEST SPECIFICATIONS

**Test Specifications for the**
**Washington Assessment of Student Learning**
**Grade 4 Mathematics**
**February, 1998**

NOTE: These are the specifications (blueprint) that guided the development of the Writing assessment based on Washington State's Essential Academic Learning Requirements.

**Test Specifications**

## I. TEST PURPOSE

The purpose of this test is to measure the level of mathematics proficiency that Washington students have achieved by the spring of the fourth grade, according to the Essential Academic Learning Requirements established by the Washington Commission on Student Learning. These Essential Academic Learning Requirements consist of four fundamental processes-solving problems, reasoning, communicating, and making connections-and the mathematical concepts that support these processes.

| | Content Strands | | Process Strands |
|---|---|---|---|
| 1 | Number Sense | 6 | Solving Problems |
| 2 | Measurement | 7 | Reasoning Logically |
| 3 | Geometric Sense | 8 | Communicating Understanding |
| 4 | Probability and Statistics | 9 | Making Connections |
| 5 | Algebraic Sense | | |

In keeping with the CSL Essential Academic Learning Requirements Technical Manual, February 26, 1997, these Essential Academic Learning Requirements-the content and process strands-are viewed as an integrated whole. Each test item will be identified as to its primary content and/or process strand it is assessing.

The following strands are intended to summarize the benchmark indicators of knowledge and skills (or Essential Academic Learning Requirements content or process examples) as identified in the mathematics section of the Essential Academic Learning Requirements Technical Manual. The benchmark indicators from the Essential Academic Learning Requirement Technical Manual are identified by numbers in parentheses after each target and are listed at the end of this document.

## II. CONTENT STRANDS AND LEARNING TARGETS[1]

### Strand 1: Number Sense (NS)

NS01 (Numbers)

Identify and illustrate whole numbers and fractions in a variety of forms and representations, using pictures, models, and symbols. (Mathematics EALR 1.1.1)

NS02 (Numeration)

Demonstrate an understanding of place value and magnitude in identifying, ordering, and comparing whole numbers and common or simple fractions. (Mathematics EALRs 1.1.1, 1.1.2)

NS03 (Computation and Conceptual Understanding of Operations)

Add, subtract, multiply, and divide whole numbers; demonstrate an understanding of whole number operations and fraction operations at the concrete level. (Mathematics EALRs 1.1.3, 1.1.4)

NS04 (Estimation)

Determine appropriateness of estimation and use estimation to predict computation results and determine reasonableness of answers. (Mathematics EALRs 1.1.6, 1.1.7)

NS05 (Number Theory)

Identify and illustrate properties of whole numbers and break down (decompose), combine, compare, pattern/sequence, and order numbers. (Mathematics EALRs 1.1.1, 1.1.2)

---

[1] Although wording throughout the learning targets may seem to indicate that all items will assess students' understanding of several concepts or procedures, the ands in the targets should be understood to mean and/or, since some items, particularly multiple-choice items, may focus only on one concept or procedure. Items will be developed for every concept/procedure.

**Strand 2: Measurement (ME)**

ME01 (Attributes and Dimensions)

Describe and compare objects and measurable attributes of objects (such as length, perimeter, area, volume or capacity, angle, weight, money, and temperature) in standard units. (Mathematics EALR 1.2.2)

ME02 (Calculation)

Select, use, and evaluate appropriate instruments, units (standard or nonstandard), and procedures for measuring time, money, length, area, volume, weight, and temperature. (Mathematics EALRs 1.2..6, 1.2.7)

ME03 (Approximation)

Use estimation to predict or determine the reasonableness of measurements and to obtain reasonable approximations. (Mathematics EALR 1.2.4)

ME04 (Systems and Precision)

Demonstrate an understanding of the appropriate uses of standard and nonstandard units of measure and the approximate nature of measurement. (Mathematics EALRs 1.2.5, 1.2.3)

**Strand 3: Geometric Sense (GS)**

GS01 (Shapes and Figures)

Identify, describe, sort, and compare geometric figures using their attributes; describe how geometric shapes and objects in the surrounding environment are related; and construct geometric figures. (Mathematics EALRs 1.3.7, 1.3.1, 1.3.2)

GS02 (Locations and Transformations)

Identify and describe the relative location of objects to one another; identify and describe the location of objects on a location grid (map, grid, number line); identify and construct simple geometric transformations using slides, flips, and turns. (Mathematics EALRs 1.3.3, 1.3.6)

GS03 (Geometric Relationships)

Identify, describe, and compare parallel, perpendicular, and intersecting lines, as well as congruent, symmetrical, and similar figures, in two-dimensional and real-world constructions. (Mathematics EALRs 1.3.4, 1.3.5)

**Strand 4: Probability and Statistics (PS)**

PS01 (Determine Probabilities)

Predict, show, and evaluate the possible outcomes and probabilities of simple experiments and activities; distinguish between certain and uncertain events; and compare predictions to experimental results. (Mathematics EALRs 1.4.1, 1.4.2, 1.4.3, 1.4.8)

PS02 (Data Collection)

Identify, describe, and evaluate methods for the effective collection of data. (Mathematics EALR 1.4.5)

PS03 (Analyze Data)

Collect,[2] organize, analyze, and display data in graphs, tables, charts, and other pictorial representations (e.g., icons); make and evaluate inferences from data and experimental results. (Mathematics EALRs 1.4.6, 1.4.9)

PS04 (Make Inferences and Predictions)

Identify, find, and use defined measures of central tendency (mean, median, mode) and other characteristics to describe a set or sets of data and sample populations. (Mathematics EALR 1.4.7)[3]

**Strand 5: Algebraic Sense (AS)**

AS01 (Patterns and Sequences)

Recognize, create, and extend patterns of objects and numbers. (Mathematics EALR 1.5.1)

AS02 (Symbols and Notation)

Identify and use appropriate symbols/notation to represent number patterns and operations, and to translate problem situations into mathematical symbols. (Mathematics EALR 1.5.3)

AS03 (Equations)

Set up and solve simple equations at the concrete or pictorial level. (Mathematics EALR 1.5.6)]

---

[2] Actual collection of data will be assessed as Classroom-Based Evidence.

[3] Understanding of the differences between samples and populations is not critical at this grade level.

## III. PROCESS STRANDS AND LEARNING TARGETS

**Strand 6: Solving Problems (SP)**

SP01 (Investigates Situations)

Use, modify, create, and evaluate strategies and approaches to conduct explorations and perform operations. (Mathematics EALR 2.3.3)

SP02 (Defines the Problem)

Formulate questions; define problems; and identify patterns, questions to be answered, missing or unnecessary data, and unknowns. (Mathematics EALRs 2.2.1, 2.2.3, 2.1.3)

SP03 (Constructs Solutions)

Collect needed information, select and use tools, use a variety of strategies, and apply concepts and procedures in constructing solutions. (Mathematics EALRs 2.1.2, 2.3.2)

**Strand 7: Reasoning Logically (RL)**

RL01 (Analyzes)

Compare and contrast information, and interpret information from a variety of sources. (Mathematics EALR 3.1.1)

RL02 (Verifies)

Identify and use models, known facts, patterns, relationships, counterexamples, and deductive and inductive reasoning to validate thinking, support arguments, and evaluate procedures and results. (Mathematics EALRs 3.3.1, 3.1.2, 3.3.4)

RL03 (Predicts)

Make inferences, predictions, and conclusions based on analysis of problem situations. (Mathematics EALRs3.2.1)

**Strand 8: Communicating Understanding (CU)**

CU01 (Planning)

Create a plan for collecting information. (Mathematics EALR 4.1.1)

CU02 (Gathering Information)

Use reading, listening, and observation skills to gather, extract, and interpret mathematical information from a variety of sources-pictures, diagrams, models, text, symbolic representations, and technology. (Mathematics EALRs 4.1.2, 4.1.3)

CU03 (Interpreting, Representing, and Sharing)

Represent, organize, and express mathematical information, understandings, and ideas using models, tables, charts, graphs, written reflections, and algebraic notation, and explain these ideas in ways appropriate to a given audience. (Mathematics EALRs 4.2.1, 4.3.1, 4.3.2)

**Strand 9: Making Connections (MC)**[4]

MC01 (Making Connections Among Concepts and Procedures)

Link conceptual and procedural understandings among the areas of number sense, measurement, geometric sense, probability and statistics, and algebraic sense. (Mathematics EALR 5.1.1)

MC02 (Equivalent Representations of Mathematical Ideas)

Use, create, and evaluate equivalent graphical, numerical, physical, algebraic, geometric, and verbal mathematical models and representations. (Mathematics EALR 5.1.2)

MC03 (Mathematics in Other Disciplines, Real Life, and the World of Work)

Identify and apply mathematical thinking, modeling, patterns, and ideas in other disciplines, real-life situations, and job-related applications. (Mathematics EALR 5.2.2, 5.3.1, 5.3.2)

*The following benchmark indicators have been identified or recommended as bases for classroom-based assessment activities:*

Mathematics EALR 1.1.5 Demonstrate ability to use mental arithmetic, pencil and paper, and calculator as appropriate; choose the appropriate strategy.

Mathematics EALR 1.5.2 Use the guess and check strategy in searching for and evaluating patterns.

---

[4] Math relations and applications within real-world situations, other disciplines, or within mathematics will permeate the test. That is, whenever possible, these relations and applications will be made. Specific items, however, will also be constructed to assess students' ability to use, identify, or construct such applications or relations.

## IV. CONTENT ORGANIZATION

The operational test forms contain 40 items, for a total of 62 points. Items are written at a reading level appropriate to a fourth-grade audience; thus, item development was aimed for an end-of-third-grade readability. Test forms include the following item types:

Multiple-choice items: The student has three or four responses to choose from-the correct answer and at least two distracters. The operational test forms contain 24 multiple-choice items.

Short-answer items (including enhanced multiple-choice[5]): The student must construct a short response, for example, write a sentence or equation; complete a table, graph, or chart; draw a picture; construct a diagram; perform a calculation. The operational test forms contain 13 short-answer items worth 2 points each.

Extended-response items: The student must construct a longer (than a short answer) response; for example, create a graph showing the appropriate data, labeled axes, and title; create and/or extend tables, diagrams, or pictures; provide a lengthy written explanation, a written explanation with number sentences, pictures, and/or diagrams, and so on. The operational test forms have 3 extended-response items worth 4 points each.

The mathematics test is designed to be administered in two sittings, each of which will be about 1 hour 15 minutes, including breaks. On each form of the test, each of the two parts will contain between 20–22 items in approximately the following proportions: 11–14 multiple-choice, 5–8 short-answer, and 1 or 2 extended-response items.

Each test form contains a variety of items so that all strands or Essential Academic Learning Requirements are addressed; thus, each form (and each of the two parts of the test) consist of a mix of items addressing content and process strands. The two parts of the test are constructed so as to separate the items on which tools (such as rulers or calculators) must not be used from the items for which tools are encouraged or possibly required. Each strand includes at least one short-answer or one extended-response item.

[5] Enhanced multiple-choice items are items in which the student selects from a list of possible responses and explains the reason(s) for choosing that response.

## V. TEST and ITEM SCORING

Each multiple-choice item is worth 1 point, each short-answer item is worth 2 points, and each extended-response item is worth 4 points. Thus, in a 40-item operational test form, 24 multiple-choice items are worth 24 points, 13 short-answer items are worth 26 points, and 3 extended-response items are worth 12 points, making a total of 62 possible points. Multiple-choice items account for about 39% of the total score points; short-answer items, 42%; and extended-response, 19%.

| Type | Number of Items | Total Points | Percent of the Total Score |
|---|---|---|---|
| Multiple-choice | 24 | 24 | 39% |
| Short-answers | 13 | 26 | 42% |
| Extended-response | 3 | 12 | 19% |
| Total | 40 | 62 | |

* No more than 3 short-answer items will be enhanced multiple-choice.

### Scoring of Open-Ended Items

Individual scoring criteria will be developed for each constructed-response item. Short-answer items will be scored on a scale of 0 to 2 points, and extended-response items will be scored on a scale of 0 to 4 points. The following scoring criteria are intended for conceptual understanding and mathematical processes. Specific scoring criteria will be developed for each item.

### Scoring Rules for Short Answer Items

**Scoring rules for items that assess concepts and procedures:**

2 A 2-point response shows complete understanding of the concept or task, as well as consistent and correct use of applicable information and/or procedures. Set-up and computations are accurate.

1 A 1-point response shows partial understanding of the concept or task. There may be minor errors in the use of applicable information and/or procedures. Set-up or computations may have minor errors.

0 A 0 point response shows little or no understanding of the concept or task.

**Scoring rules for items that assess communicating understanding:**

2 A 2-point response shows understanding of how to effectively and appropriately interpret, organize, and/or represent mathematical information relevant to the concept.

1 A 1-point response shows some understanding of how to interpret, organize, and/or represent mathematical information relevant to the concept; however, the response is not complete or effectively presented.

0 A 0 point response shows little or no understanding of how to interpret, organize and/or represent mathematical information relevant to the concept.

**Scoring rules for items that assess <u>solving problems</u>:**

2 A 2-point response shows thorough investigation, clear understanding of the problem, and/or effective and viable solution.

1 A 1-point response shows partial investigation and/or understanding of the problem, and/or a partially complete or partially accurate solution.

0 A 0-point response shows very little or no investigation and/or understanding of the problem, and/or no visible solution; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

**Scoring rules for items that assess <u>reasoning logically</u>:**

2 A 2-point response shows effective reasoning through a complete analysis or thorough interpretation, supported predictions, and/or verification.

1 A 1-point response shows somewhat flawed reasoning either through incomplete analysis or interpretation, prediction that lacks support, or inadequate verification.

0 A 0-point response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

## Scoring Rules for Short Answer Items (Cont.)

**Scoring rules for items that assess making connections:**

2 A 2-point response makes clear and effective connections within and/or between conceptual or procedural areas.

1 A 1-point response makes vague or partially accurate connections within and/or between conceptual or procedural areas.

0 A 0-point response makes little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?)

## Scoring Rules for Extended Response Items

**Scoring rules for items that assess solving problems:**

4 points -- Meets all relevant criteria

- Thoroughly investigates the situation
- Uses all applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs elegant, efficient, valid solution using applicable tools and workable strategies

3 points -- Meets all or most relevant criteria

- Investigates the situation
- Uses most applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs viable/acceptable solution using applicable tools and workable strategies

2 points -- Meets some relevant criteria

- Investigates the situation, but may omit issues or information
- Uses some applicable information related tot he problem
- Uses some applicable mathematical concepts and procedures
- Constructs solution using applicable tools and workable strategies, solution may not completely address all issues or strategies may have flaws

1 point -- Meets few relevant criteria

- Attempts to investigate the situation
- Uses some applicable information related to the problem
- Uses few applicable mathematical concepts and procedures
- Attempts solution, however, mostly incomplete or not effective

0-points--Student's response provides no evidence of problem-solving skills or shows very little or no understanding of the task; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## Scoring Rules for Extended Response Items (Cont.)

**Scoring rules for items that assess communicating understanding:**

4 points -- Meets all relevant criteria

- Gathers all applicable information form appropriate sources
- Demonstrates interpretations and understandings in a clear, systematic, and organized manner
- Represents mathematical information and ideas in an effective format for the task, situation, and audience

3 points -- Meets most relevant criteria

- Gather applicable information from appropriate sources
- Demonstrates interpretations and understandings in a clear and organized manner
- Represents mathematical information and ideas in an expected format for the task, situation, and audience

2 points -- Meets some relevant criteria

- Gathers information from appropriate sources
- Demonstrates interpretation and understandings in an understandable manner
- Represents mathematical information in an acceptable format for the task, situation, and audiences

1 point -- Meets few relevant criteria

- Gathers little information from appropriate sources
- Demonstrates interpretations and understandings in a manner that may be disorganized or difficult to understand
- Represents mathematical information and ideas in a format that may be inappropriate for the task, situation, and audience.

0-points--Student's response shows little or no understanding of how to interpret, organize or represent mathematical information relevant to the concept; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## Scoring Rules for Extended Response Items (Cont.)

**Scoring rules for items that assess reasoning logically:**

4 points -- Meets all relevant criteria

- Makes insightful interpretations, comparisons, or contrasts of information form sources
- Effectively uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes insightful conjectures and inferences, if asked
- Systematically and successfully evaluates effectiveness of procedures and results, if asked
- Gives comprehensive support for arguments and results

3 points -- Meets most relevant criteria

- Makes thoughtful interpretations, comparisons, or contrasts of information form sources
- Uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes expected conjectures and inferences, if asked
- Successfully evaluates effectiveness of procedures and results, if asked
- Gives substantial support for arguments and results

2 points -- Meets some relevant criteria

- Makes routine interpretations, comparisons, or contrasts of information form sources
- Includes examples, models, facts, patterns, or relationships to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Partially evaluates effectiveness of procedures and results, if asked
- Gives partial support for arguments and results

1 point -- Meets few relevant criteria

- Makes superficial interpretations, comparisons, or contrasts of information form sources
- Examples, models, facts, patterns, or relationships may not be included to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Attends to wrong information and/or persists with faulty strategy when evaluating effectiveness of procedures and results
- Support for arguments and results may not be included

0-points--Student's response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## Scoring Rules for Extended Response Items (Cont.)

**Scoring rules for items that assess making connections:**

4 points -- Meets all relevant criteria

- Shows a thorough understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in a clear and insightful manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in a clear and insightful manner

3 points -- Meets most relevant criteria

- Shows a general understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in an obvious/expected manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in an obvious/expected manner

2 points -- Meets some relevant criteria

- Shows a partial understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations

1 point -- Meets few relevant criteria

- Shows a little understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, mathematical patterns and concepts in other disciplines AND/OR
- Identifies applies mathematical patterns and concepts in real-life situations

0-points--Student's response makes very little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## VI. REPORTING SCHEME AND ITEM DISTRIBUTION

Student performance will be reported on each of the Essential Academic Learning Requirements in the fourth-grade mathematics test. A single, comprehensive total math score is reported based on the performance standards established by the standard-setting committee. In addition, the five content and four process strands are reported as strengths or weaknesses. All Essential Academic Learning Requirements, or content and process strands, are addressed in each test form. Each content strand includes one short-answer item; all extended-response items are constructed for the process strands only.

**Mathematics Test: Overall item distribution for the operational test forms is as follows:**

| Test Strands | Number of Items (approximate range) | Number of Points (approximate range) |
|---|---|---|
| Number Sense | 4-8 | 5-9 |
| Measurement Concepts | 4-7 | 5-8 |
| Geometric Sense | 3-6 | 5-7 |
| Probability and Statistics Procedures | 4-7 | 5-8 |
| Algebraic Sense | 3-6 | 4-7 |
| Solving Problems | 2-6 | 6-12 |
| Reasoning Logically | 2-6 | 6-12 |
| Communicating Understanding | 2-6 | 6-12 |
| Making Connections | 2-5 | 4-12 |

Content total = approx. 34 pts.
Process total= approx. 28 pts.
The nine strands average about 7 points per strand, although the process strands may have slightly more points from one form to another.

**Mathematics Test: Distribution according to item type within strands**

| Test Strands | Multiple Choice | Short Answer | Extended Response | Total Number of Items |
|---|---|---|---|---|
| Number Sense | 2-6 | 1-2 | 0 | 4-8 |
| Measurement Concepts | 2-6 | 1-2 | 0 | 4-8 |
| Geometric Sense | 3-6 | 1-2 | 0 | 4-8 |
| Probability and Statistics Procedures | 3-6 | 1-2 | 0 | 4-8 |
| Algebraic Sense | 3-6 | 1-2 | 0 | 4-8 |
| Solving Problems | 0-2 | 1-2 | 1-2 | 2-6 |
| Reasoning Logically | 0-2 | 1-4 | 0-1 | 2-5 |
| Communicating Understanding | 0-2 | 1-4 | 0-1 | 2-5 |
| Making Connections | 0-2 | 1-4 | 0 | 2-5 |
| Total Number of Items | 24 | 13 | 3 | 40 |
| Total Number of Points | 24 | 26 | 12 | 62 |

## VII. GENERAL CONSIDERATIONS

- The test is designed to be administered in two separate sessions, each of which will be about 1 hour and 15 minutes long, including breaks that students may take. The test is not a specifically timed test, but total testing time for a standard administration should be about 2 1/2 hours.
- Each multiple-choice item has three or four responses-the correct answer and at least two but sometimes three distracters. Distracters were developed based on the types of errors most commonly made by students. Correct responses are approximately equally distributed among As, Bs, and Cs (and Ds). D distracters were added only if the increase in reading load was minimal and if the extra answer choice was reasonable and potentially attractive to students.
- In the item pool, item codes accompany each item and provide information regarding the content or process strand addressed, learning target addressed, item format, correct answer key (as appropriate), and any graphics associated with the item. The following abbreviations were used to indicate content and process strands in the item codes:
- Each test form contains items assessing learning targets from all content and process strands.
- Test items that assess each learning target will not be limited to one particular type of response format. However, extended-response formats are reserved for those items that assess learning targets in the mathematics process strands.

- Test questions attempt to focus on content that is "real-world" and that fourth-grade students can relate to. Test items are worded precisely and clearly. The better focused an item, the more reliable and fair it is certain to be, and the more likely all students will understand what is required of them.
- Scoring criteria for all constructed-response items focus on the clear communication of mathematical ideas, information, and solutions. The conventions of writing (sentence structure, word choice, usage, grammar, spelling, and mechanics) are disregarded, as long as they do not interfere with communicating the response.
- All items were reviewed to eliminate language or content that was biased, offensive, or disadvantageous to a particular group of students. No items that display or imply unfair representations of gender, race, persons with disabilities, or cultural or religious groups are included.

VIII. NOTATIONAL CONSIDERATIONS FOR GRADE 4

- In the item stems, numbers (other than years) having more than three digits to the left of the decimal point include commas to group digits in the usual manner (e.g., 135,000).
- Units are given when appropriate. Standard abbreviations may be used (e.g., cm or ft). However, the unit is spelled out if any confusion is reasonably possible.
- Variables are always italicized. The italicized variable x is not used to avoid confusion with the multiplication sign.
- The symbols • and * are not used as multiplication signs in items to test students' multiplication abilities. Only the symbol x is used as the multiplication sign.
- Fractions have horizontal lines separating numerator and denominator. Only common and simple fractions are used to test learning targets at 4th grade level.
- For operations such as addition and subtraction, as well as for comparisons and ordering, only common and simple fractions are used, and then generally with pictorial representations.
- Grids are used in test items that involve finding the area of a geometric figure. Illustrations are used in test items that involve finding the volume.
- Decimals are used only when expressing monetary units; expressions for monetary units use the cents sign (not decimals) for items less than a dollar-e.g., 25¢. [But do use dollar sign and decimals in mixed cases-e.g., $1.25 in dollars and cents.]

**IX. CHARACTERISTICS OF ITEMS AND ITEM STEMS OR FOILS**

**General Characteristics**

- To the extent possible, reading will be kept to a minimum to help make items clear and precise.
- Character names on each form are representative of the ethnic diversity of Washington students. The names are generally short and simple to read.

- To the extent possible, no stimulus, stem, or response for an item will serve as a clue to the correct response for another item.

- All items and stimulus materials will avoid language that shows bias, offends, or disadvantages a particular group of students. That is, items will not display unfair representations of gender, race, persons with disabilities, or cultural or religious groups.

- Items and stimulus materials in each form are balanced by gender and are gender-neutral for active/passive roles.

**Characteristics of Stimulus Materials**

- A stimulus that gives information might precede a question or a set of questions. A stimulus consists of brief written material and/or a graphic, such as a simple diagram, graph, chart, table, or drawing.

- The stimulus for an item is always factually correct and has a readability level targeted for an end-of-fifth-grade. Stimuli adapted specifically for the test. Test items focus on what is essential and consequential in the stimulus and minimize the impact of, or need for, outside (prior) knowledge.

- Graphs, tables, or figures are always clearly associated with their intended items. Graphics will appear either on the same page as the stimulus or on the facing page. If there is any reasonable chance for confusion, page references direct students to look at the appropriate graphic.

- Pictorial representations will be realistic and authentic for fourth graders.

- On items for which manipulatives and/or tools are encouraged or required, students may be given the opportunity to use any punch-out or overlay manipulatives provided, or may use those classroom manipulatives or tools with which they are most familiar/comfortable, as long as nothing about the tools would introduce bias into results. Tools include the following: rulers, geoboards, and pattern blocks.

**Item Characteristics**

- Each item begins with a stem that asks a question or gives a prompt. A stem usually asks a direct question or gives clear directions. It seldom uses an incomplete sentence, is worded negatively, or asks for a "best" answer.

- Test items are independent in the sense that the answer for any test item does not depend on knowing the correct answer to another item, so items are not "linked." Note: Linkage will be avoided among different items, not necessarily among parts within a single item. For instance, an enhanced multiple-choice may ask students to explain their reason for selecting a particular response. This is not linking between items.

- When appropriate, several items may center around a particular stimulus, graph, chart, or scenario, in which case, these items will generally appear on the same page or facing page from the stimulus.

- All items will clearly indicate what is expected in a response to help students focus their responses. That is, items will clearly state the criteria by which the response will be evaluated, so that students understand what they are expected to do (e.g., create a table, provide a written explanation, calculate a solution, etc.). General directions that allow the student more freedom in response format may read as follows, "Using words, numbers, and/or pictures, show or explain your thinking". In such cases, any of these response modes is acceptable as long as it is complete and responsive to the item stem.

- Items testing application and problem solving will involve understandable, realistic situations to which as many fourth graders as possible can relate.

- All multiple-choice items-key and distracters-are to be similar in length and in syntax; students should not be able to rule out a wrong answer or identify a correct response simply by virtue of its looking or sounding different.

- Correct responses will be approximately equally distributed among As, Bs, and Cs and Ds. Response choices like "Both of the above," "All of the above," "None of the above," and "Neither of the above" are not be used. The use of the word not is generally avoided in item stems.

- The most likely incorrect answers are generally included as distracters for multiple-choice items.

- Distracters were created so that students must think their way to the correct answer rather than simply identify incorrect responses by virtue of a distracter's obviously inappropriate nature. Distracters should always be plausible (but of course incorrect) in the context of the item stem. The responses or distracters will be arranged in a logical order, i.e., numerical or chronological order or according to the length of the distracters.

- Care will be taken not to use items for which wrong methods yield the correct response. For example, "Simplify the fraction 64/16" is a poor item, since the correct response can be obtained by canceling the two sixes.

- If a question is stated in terms of one measurement system, all response options should be given in terms of the same measurement system. Units do not have to be included in the stem, but they should appear in every distracter or response when appropriate.

## X. COMPONENTS AND BENCHMARKS

The following benchmark indicators come from Essential Academic Learning Requirements Technical Manual, Washington State Commission on Student Learning, February 24, 1997

### 1. Student Understands and Applies the Concepts and Procedures of Mathematics

1.1 Understand and Apply Concepts and Procedures from Number Sense

1.1.1 uses objects, pictures, or symbols to demonstrate understanding of whole and fractional numbers, place value in whole numbers, and properties of the whole number system

1.1.2 identify, compare, and order whole numbers and simple fractions

1.1.3 show understanding of whole number operations (addition, subtraction, multiplication and division) using blocks, sticks, beans, etc.

1.1.4 add, subtract, multiply and divide whole numbers

1.1.5 uses mental arithmetic, pencil and paper, or calculator as appropriate to the task involving whole numbers

1.1.6 identifies situations involving whole numbers in which estimation is useful

1.1.7 uses estimation to predict computation results and to determine the reasonableness of answers, for example, estimating a grocery bill

1.2 Understand and Apply Concepts and Procedures from Measurement

1.2.1 understands concepts of perimeter, area, and volume

1.2.2 use directly measurable attributes such as length, perimeter, area [illegible] angle, weight/mass, money, and temperature to describe and [illegible]

1.2.3 understands that measurement is approximate

1.2.4 knows how to estimate to predict and to determine when [illegible] reasonable, for example, estimating the length of the playground by [illegible]

1.2.5 understands the benefits of using standard units of measure[illegible] measuring length, area, and volume

1.2.6 knows appropriate units of measure for time, money, length, area, volume, mass, and temperature

1.2.7 uses appropriate tools for measuring time, money, length, area, volume, mass, and temperature

1.3 Understand and Apply Concepts and Procedures from Geometric Sense

1.3.1 use shape and size to identify, name, and sort geometric shapes

1.3.2 recognize geometric shapes in the surrounding environment, for example, identify rectangles within windows

1.3.3 describes the relative location of objects relative to each other on grids or maps

1.3.4 understands concepts of parallel and perpendicular

1.3.5 understands concepts of symmetry, congruence, and similarity

1.3.6 understands and constructs simple geometric transformations using slides, flips, and turns

1.3.7 constructs simple shapes using appropriate tools such as a straightedge or a ruler

1.4 Understand and Apply Concepts and Procedures from Probability and Statistics

1.4.1 understands the difference between certain and uncertain events

1.4.2 knows how to list all possible outcomes of simple experiments

1.4.3 understands and uses experiments to investigate uncertain events

1.4.4 knows that data can be represented in different forms such as tabulations of events, objects, or occurrences

1.4.5 can collect data in an organized way

1.4.6 organize and display data in numerical and graphical forms such as tables, charts, pictographs, and bar graphs

1.4.7 use different measures of central tendency such as "most often" and "middle" describing a set of data

1.4.8 predict outcomes of simple activities and compares the predictions to experimental results

1.4.9 understands and makes inferences based on experimental results using coins, number cubes, spinners, etc.

1.5 Understand and Apply Concepts and Procedures from Algebraic Sense

1.5.1 recognize, create and extend patterns of objects and numbers using a variety of materials such as beans, toothpicks, pattern blocks, calculator, cubes, or colored tiles

1.5.2 understands the use of guess and check in the search for patterns

1.5.3 represent number patterns symbolically, for example, using tiles, boxes, or numbers

1.5.4 use standard notation in reading and writing open sentences, for example, 3 x· = 18

1.5.5 evaluate simple expressions using blocks, sticks, beans, pictures, etc.

1.5.6 solve simple equations using blocks, sticks, beans, pictures, etc.

**2. Student Uses Mathematics to Define and Solve Problems**

2.1 Investigate Situations

2.1.1 search for patterns in simple situations

2.1.2 uses a variety of strategies and approaches

2.1.3 recognizes when information is missing or extraneous

2.1.4 recognizes when an approach is unproductive and tries a new approach

2.2 Formulate Questions and Define the Problem

2.2.1 identifies questions to be answered in familiar situations

2.2.2 defines problems in familiar situations

2.2.3 identify the unknowns in familiar situations

2.3 Construct Solutions

2.3.1 organizes relevant information

2.3.2 select and use appropriate mathematical tools

2.3.3 apply appropriate methods, operations, and processes to construct a solution

**3. Student Uses Mathematical Reasoning**

3.1 Analyze Information

3.1.1 interpret and compare information in familiar situations

3.1.2 validate thinking using models, known facts, patterns, and relationships

3.2 Predict Results and Make Inferences

3.2.1 makes conjectures and inferences based on analysis of familiar problem situations

3.3 Draw Conclusions and Verify Results

3.3.1 tests conjectures by finding examples to support or contradict them

3.3.2 supports arguments and justify results based on own experiences

3.3.3 checks for reasonableness of results

3.3.4 reflects on and evaluates procedures and results in familiar situations

**4. Student Communicates Knowledge and Understanding in Both Everyday and Mathematical Language**

4.1 Gather Information

4.1.1 follow a plan for collecting information

4.1.2 uses reading, listening, and observation skills to access and extract mathematical information from a variety of sources such as pictures, diagrams, physical models, classmates, oral narratives, and symbolic representations

4.1.3 use available technology to browse and retrieve mathematical information from a variety of sources

4.2 Organize and Interpret Information

4.2.1 organize and clarify mathematical information in at least one way reflecting, verbalizing, discussing, or writing

4.3 Represent and Share Information

4.3.1 express ideas using mathematical language and notation such as physical or pictorial models, tables, charts, graphs, or symbols

4.3.2 expresses mathematical ideas to familiar people in everyday language

**5. Student Understands How Mathematical Ideas Connect Within Mathematics, to Other Subject Areas, and to Real-life Situations**

5.1 Relate Concepts and Procedures Within Mathematics

5.1.1 connect conceptual and procedural understandings among familiar mathematical content areas

5.1.2 recognize equivalent mathematical models and representations in familiar situations

5.2 Relate Mathematical Concepts and Procedures to Other Disciplines

5.2.1 recognize mathematical patterns and ideas in familiar situations in other disciplines

5.2.2 uses mathematical thinking and modeling in familiar situations in other disciplines

5.2.3 describes examples of contributions to the development of mathematics such as the contributions of women, men, and different cultures

5.3 Relate Mathematical Concepts and Procedures to Real-life Situations

5.3.1 give examples of how mathematics is used in everyday life

5.3.2 identify how mathematics is used in career settings

# APPENDIX C

# WASHINGTON ASSESSMENT OF STUDENT LEARNING

## GRADE 4

## READING AND LISTENING

## TEST SPECIFICATIONS

**Test Specifications for the**
**Washington Assessment of Student Learning**
**Grade 4 Reading and Listening**
**February, 1998**

NOTE: These are the specifications (blueprint) that guided the development of the Writing assessment based on Washington State's Essential Academic Learning Requirements.

## I. PURPOSE

The purpose of this test is to measure Washington fourth-grade students' level of proficiency in the Essential Academic Learning Requirements in reading. The reading test contains literary, informational, and task-oriented reading selections. All reading selections, ranging up to 600 words and written at a difficulty level appropriate for fourth grade students, are accompanied by test items that assess proficiency in the components of the Essential Academic Learning Requirements in reading. Test items are of the following types:

- Multiple-choice: Student chooses from three responses provided.
- Short-answer: Student constructs short response-phrase(s) or sentence(s).
- Extended-response: Student constructs longer, more sustained response sentences or paragraph(s).

Each form of the reading and listening tests attempt to cover all identified Learning Targets, but this may not always be practical; not every text allows every type of question to be asked. A single, comprehensive total reading score is reported based on the performance standards established by a standard-setting committee. A single, comprehensive total listening score is reported based on the performance standards established by a standard-setting committee. The reading test also offers two subscale reports (reported as strengths or weaknesses) in Reading for Literary Experience and two subscale reports in Reading to Learn New Information and to Perform a Task. The first subscale report for each type of text reflects students' comprehension of important ideas and details, and the second reflects students' ability to analyze, interpret, and think critically about what they have read.

## II. LEARNING TARGETS (STRANDS)

**Reading for Literary Experience ([RL] Reading EALR 3.3):**

RL01 Target-Comprehends important ideas and details (Reading EALR 2.1)

Given a literary text to read silently, learners respond to items in which they:

- Demonstrate understanding of theme or message and supportive details (Reading EALR 2.1.2)
- Summarize with evidence from the reading (Reading EALR 2.1.2)
- Make inferences or predictions based on the reading (Reading EARL 2.1.4)
- Interpret vocabulary critical to the meaning of the text (Reading EALR 1.2.1)
- Order steps, sequence, and/or parts from the reading (Reading EALRs 1.4.2, 2.2.2)

RL02-Analyzes, interprets, and thinks critically (Reading EALRs 2.2, 2.3)

*Analyzes and interprets*

Given a literary text to read silently, learners respond to items in which they:

- Demonstrate understanding of literary elements (genres; story elements such as plot, character, setting; stylistic devices) and graphic elements/illustrations (Reading EALR 1.4.3)
- Compare and contrast elements of text (Reading EALR 2.2.1)
- Make connections within and among texts (Reading EALR 2.2.1)

*Thinks critically*

Given a literary text to read silently, learners respond to items in which they:

- Analyze author's purpose and evaluate effectiveness for different audiences (Reading EALR 2.3.2)
- Extend information beyond text—apply information, give a response to reading, express insight gained from reading (Reading EALR 2.3.3)

**Reading to Learn New Information and to Perform Tasks ([RI] Reading EALRs 3.1 and 3.2)**

RI01 Comprehends important ideas and details

Given an informational or task-oriented text to read silently, learners respond to items in which they:

- Demonstrate understanding of theme or message and supportive details (Reading EALR 2.1.2)
- Summarize with evidence from the reading (Reading EALR 2.1.2)
- Make inferences or predictions based on the reading (Reading EARL 2.1.4)
- Interpret vocabulary critical to the meaning of the text (Reading EALR 1.2.1)
- Order steps, sequence, and/or parts from the reading (Reading EALRs 1.4.2, 2.2.2)

RI02 Analyzes, interprets, and thinks critically

*Analyzes and interprets*

Given an informational or task-oriented text to read silently, learners respond to items in which they:

- Demonstrate understanding of text features (titles, headings, and other information divisions, table of contents, captions) and graphic features Reading EALR 1.5.2)
- Compare and contrast elements of text (Reading EALR 2.2.1)
- Make connections within and among texts (Reading EALR 2.2.1)

*Thinks critically*

Given an informational or task-oriented text to read silently, learners respond to items in which they:

- Analyze author's purpose (including distinguishing between fact and opinion) and evaluate effectiveness for different audiences (Reading EALRs 2.3.1, 2.3.2)
- Extend information beyond text—apply information, give a response to reading, express insight gained from reading (Reading EALR 2.3.3)

**Listening to Gain Understanding ([LI] Communication EALRs 1.2.4, 1.3.2)**

LI01 Listening to gain understanding

Given an orally presented text, learners respond to items in which they:

- Identify and explain main ideas, details, facts vs. opinions, and meaning (Communication EALR 1.2.4)
- Paraphrase information (Communication EALR 1.3.2)

## III. CONTENT ORGANIZATION

### General Characteristics

- It is not possible to measure every Learning Target on every form of the test. However, Learning Targets from each strand are tested on each form.
- The material presented is balanced, culturally diverse, well written, and of interest to 4th-grade students. The passages and items are fairly presented in order to gain a true picture of students' reading skills.
- Across all forms, a balance of gender and active/passive roles by gender is maintained.
- Character names on each form are representative of the ethnic diversity of Washington students.
- No resource materials may be used by students during the testing of reading.
- Responses are scored with emphasis on communication of ideas. Conventions of writing (sentence structure, word choice, usage, grammar, and mechanics) are generally disregarded unless they substantially interfere with communication.

Estimated Time for the Reading Test: 75 minutes, divided into two sessions with a break in between.

Estimated Time for Listening Test: 25 minutes.

### Characteristics of Reading and Listening Passages

Reading passages used in the test are drawn from published sources and include poetry, essays, short stories, novel/book excerpts, plays, pamphlets, and newspaper and magazine articles. Task-oriented texts consist of schedules, recipes, instructions, and other such pieces likely to be within the experience of a fourth-grade student. As appropriate, passages utilize illustrations and other graphic features. Each assessment form contains one selection that is made up of two short passages, e.g., a poem and a short piece of fiction, or a set of directions and a short piece of informational text. This pairing allowed construction of items that call for students to make connections among texts.

All passages are reviewed to eliminate cultural or other forms of bias that might disadvantage any group (or groups) of students. The passages avoid subject matter that might prompt emotional distress on the part of some students. It was critical that the reading texts used be well written, of interest to fourth-grade students, and, in all appropriate cases, factually correct. Reading test passages also reflect Washington's cultural diversity, and as they are presented they possess structural integrity that allows them to be self-contained. Permission to use selections from copyrighted material were obtained as necessary. The reading difficulty of the passages was validated using traditional readability formulas and teacher judgment. The overall suitability of each passage was judged by Washington's reading content committee.

**Characteristics of Test Items**

- Items deal with issues and details that are of consequence in the text and central to students' understanding and interpretation of the text.
- Test items are varied and address as many Learning Targets as the passages allow.
- To the greatest extent possible, no item or response choice "clues" the answer to any other item.
- The Learning Target assessed has been specified for each item.
- Items are reviewed to eliminate unfair representations of gender, race, individuals with disabilities, or cultural or religious groups.
- Test items are tied closely and particularly to the passage from which they derive, so that the impact of outside (prior) knowledge, while never wholly avoidable, was minimized.
- Each reading test form contains 18-22 multiple-choice items and each of these has one correct answer and two distracters.
- Each listening test form contains 6-8 multiple choice items and each of these has one correct answer and two distractors.
- Each multiple-choice item contains a question (or incomplete statement) and three answer (or completion) options, only one of which is correct. Correct answers are distributed as evenly as possible among A's, B's, and C's.
- The three choices are approximately the same length, have the same format, and are syntactically and semantically parallel; students should not be able to rule out a wrong answer or identify a correct response simply by virtue of its looking or sounding different.
- Distracters adopt the language and sense of the material in the selection so that students must think their way to the correct answer rather than simply identify incorrect responses by virtue of a distracter's obviously inappropriate nature.
- Distracters should always be plausible (but of course incorrect) in the context of the selection.
- All items must clearly indicate what is expected in a response and must help students focus their response.
- Constructed-response items are of two types: short-answer and extended-response.
- Each short-answer or extended-response item gives clear indications of what is required of students; an item-specific scoring guide was developed for each short-answer and extended- response item, and information from the pilot was used to refine these scoring guides for use with the final forms of the test. Scoring guides follow a "focused holistic" model in which the score for the response is based on overall quality, but also results from focusing on several important features of the student's performance.

a. Short-answer items are scored with a 3-level scoring guide (0-2) in which students may receive full credit, partial credit, or no credit.

b. Extended-response items are scored with a 5-level scoring guide (0-4); the levels may be summarized as Extensive, Essential, Partial, Minimal, and Unsatisfactory.

- The reading test form contains 7-9 short-answer items, each of which requires students to construct a short response, defined as phrase(s) or sentence(s), and focusing on one Strand.
- The listening test form contains 2 short-answer items, each of which requires students to construct a short response, defined as phrase(s) or sentence(s), and focusing on one Strand.
- Each reading test form also contains 2 extended-response items, each of which requires students to construct a longer, more sustained response, defined as sentences or paragraph(s), and focusing on one Strand.
- Extended-response items generally require longer more detailed responses providing more evidence, information, or examples.
- The two types of constructed-response items may also be differentiated by the number of lines available for the response.
- Order of presentation of item types is dictated by logic.
- When an item has two parts, they appear separately, with lines following each part. This encourages students to notice and answer both parts.
- With paired passages, items for each follow each passage. The items requiring comparison between the passages appear last, after both passages. There will always be more than one item that compares the two passages, and preferably more than two.
- Care is taken that all items avoid language that shows bias or is otherwise likely to be offensive to or to disadvantage a particular group of students.
- Items are worded precisely and clearly. The better focused an item, the more reliable and fair it is certain to be, and the more likely all students will understand in the same way what is required of them.

## IV. TEST AND ITEM SCORING

Each multiple-choice item is worth 1 point, each short-answer item is worth 2 points, and each extended-response item is worth 4 points.

**Reading Test: Approximate distribution of score points by item type**

| Type | Number of Items | Total Points | Percent of the Total Score |
|---|---|---|---|
| Multiple-choice | 18-22 | 18-22 | 45-46% |
| Short-answers | 7-9 | 14-18 | 35-37% |
| Extended-response | 2 | 8 | 17-20% |
| Total | 27-33 | 40-48 | |

**Listening Test: Approximate distribution of score points by item type**

| Type | Number of Items | Total Points | Percent of the Total Score |
|---|---|---|---|
| Multiple-choice | 6-8 | 6-8 | 60-67% |
| Short-answers | 2 | 4 | 33-40% |
| Total | 8-10 | 10-12 | |

### SCORING OF OPEN-ENDED ITEMS

Individual scoring criteria will be developed for each constructed-response item. Short-answer items will be scored on a scale of 0 to 2 points, and extended-response items will be scored on a scale of 0 to 4 points. The following scoring criteria are used to assess basic comprehension of main ideas and details and analysis, interpretation, and critical thinking about text. Specific scoring criteria will be developed for each item based on these generic rules.

## Scoring Rules for Short Answer Items

**Scoring rules for items that assess <u>main ideas and details</u>:**

**2** A two point response:
- shows thorough comprehension of main idea and important details
- uses ample, relevant information from text(s) to support responses

**1** A one point response:
- shows partial comprehension of main idea and important details (may grasp main idea but show difficulty distinguishing between important and unimportant details; may miss part of fundamental who/what/where/when/why)
- attempts to use information from text(s) to support responses; support may be limited or irrelevant

**0** A 0 point response shows little or no understanding of the passage main ideas and details.

**Scoring rules for that assess <u>analysis, interpretation, and critical hinking about text</u>:**

**2** A two point response:
- analyzes appropriate information and/or makes thoughtful connections between whole texts/parts of texts
- develops thoughtful interpretations of text
- uses sufficient, relevant evidence from text(s) to support claims

**1** A one point response:
- analyzes limited information and/or makes superficial connections between whole texts/parts of texts
- develops conventional or simplistic interpretations of text
- attempts to use evidence from text(s) to support claims; support may be limited or irrelevant

**0** A 0 point response shows little or no understanding of the passage main ideas and details.

**Scoring rules for items that assess <u>summarizing and paraphrasing main ideas</u>:**

**2** A two point response shows thorough comprehension of main ideas

**1** A one point response shows partial comprehension of main ideas

**0** A 0 point response shows little or no understanding of the passage main ideas and details.

## Scoring Rules for Extended Response Items

**Scoring rules for items that assess analysis, interpretation, and thinking about text:**

4 Points: Meets all relevant criteria

- thoroughly analyzes appropriate information and/or makes insightful connections between whole texts/parts of texts
- develops insightful interpretations of text
- uses ample, relevant evidence from text(s) to support claims

3 Points: Meets or most all relevant criteria

- analyzes appropriate information and/or makes thoughtful connections between whole texts/parts of texts
- develops thoughtful interpretations of text
- uses sufficient, relevant evidence from text(s) to support claims

2 Points: Meets some relevant criteria

- analyzes limited information and/or makes superficial connections between whole texts/parts of texts
- develops conventional or simplistic interpretations of text
- attempts to use evidence from text(s) to support claims; support may be limited or irrelevant

1 Point: Meets few relevant criteria

- shows difficulty analyzing information and/or makes weak connections between whole texts/parts of texts
- may not develop beyond literal interpretation of text
- uses little or no evidence to support claims

## DISTRIBUTION OF READING SELECTIONS, TEST ITEMS, AND SCORE POINTS

Each reading test form includes two or three literary selections, generating approximately half the total test points, and one or two informational selections and one or two task-oriented selections, generating approximately half the total test points. In addition,

- One selection in each form of the test consists of two short passages, e.g., a poem and a short piece of fiction, or a set of directions and a short piece of informational text. This pairing allowed construction of items that call for students to make connections among texts.
- Many of the selections are short, i.e., 200-300 words.
- One selection on a form may be longer (as long as 600 words) to allow for development of items that go with more extended text.
- The reading selections together total about 1500 words.

- Total number of multiple-choice items does not exceed 22.
- Total number of short-answer items does not exceed 9.
- Total number of extended-response items does not exceed 2.

**Reading Test: Item distribution by text type and strand**

| Text types/ Strands | Number of Reading Selections | Number of Words Per Passage | Number of Multiple-Choice Items | Number of Short Answer Items | Number of Extended Response Items |
|---|---|---|---|---|---|
| **Literary** | 2-3 | up to 750 | 9-12 | 3-5 | 1 |
| Comprehends important ideas and details | | | 3-8 | 1-2 | 0 |
| Analyzes, interprets, and thinks critically | | | 4-8 | 2-4 | 1 |
| **Information and Completing a Task** | 1-2 | up to 750 | 9-12 | 3-5 | 1 |
| Comprehends important ideas and details | | | 3-8 | 1-2 | 0 |
| Analyzes, interprets, and thinks critically | | | 4-8 | 2-4 | 1 |
| Total | 4-5 | up to 1500 | 18-22 | 7-9 | 2 |

**Listening Test: Item distribution by text type and strand**

| Learning Targets | Number of Reading Selections | Number of Words Per Passage | Number of Multiple-Choice Items | Number of Short Answer Items |
|---|---|---|---|---|
| Listening for important ideas and details | 1 | up to 200 | 6-8 | 2 |
| | | | 6-8 | 0 |
| Paraphrases and summarizes main ideas | | | 0 | 2 |
| Total | 1 | up to 200 | 6-8 | 2 |

# APPENDIX D

## General Scoring Rules for the Washington Assessment of Student Learning

**Listening**

**Reading**

**Mathematics**

**Writing**

## SCORING OF OPEN-ENDED LISTENING ITEMS

Individual scoring criteria were developed for each constructed-response item. Short-answer listening items were scored on a scale of 0 to 2 points. The following scoring criteria were used to guide item writers in their development of item specific scoring criteria. This helped to ensure that the item scoring criteria were clearly focused on summarizing information and paraphrasing main ideas.

### Scoring Criteria for Short Answer Listening Items

**SUMMARIZING AND PARAPHRASING MAIN IDEAS:**

**2** A two point response shows thorough comprehension of main ideas or an accurate summary of events.

**1** A one point response shows partial comprehension of main ideas or a partially accurate summary of events.

**0** A 0 point response shows little or no understanding of the passage main ideas or events.

## SCORING OF OPEN-ENDED READING ITEMS

Individual scoring criteria were developed for each constructed-response item. Short-answer items were scored on a scale of 0 to 2 points, and extended-response items were scored on a scale of 0 to 4 points. The following scoring criteria were used to guide item writers in their development of item specific scoring criteria. This helped to ensure that the item scoring criteria were clearly focused on the appropriate dimension of reading performance: basic comprehension of main ideas and details and analysis OR analysis, interpretation, and critical thinking about text.

### Scoring Criteria for Short Answer Reading Items

**MAIN IDEAS AND DETAILS:**

**2** A two point response:
- shows thorough comprehension of main idea and important details
- uses ample, relevant information from text(s) to support responses

**1** A one point response:
- shows partial comprehension of main idea and important details (may grasp main idea but show difficulty distinguishing between important and unimportant details; may miss part of fundamental who/what/where/when/why)
- attempts to use information from text(s) to support responses; support may be limited or irrelevant

**0** A 0 point response shows little or no understanding of the passage main ideas and details.

**ANALYSIS, INTERPRETATION, AND CRITICAL THINKING ABOUT TEXT:**

**2** A two point response:
- analyzes appropriate information and/or makes thoughtful connections between whole texts/parts of texts
- develops thoughtful interpretations of text
- uses sufficient, relevant evidence from text(s) to support claims

**1** A one point response:
- analyzes limited information and/or makes superficial connections between whole texts/parts of texts
- develops conventional or simplistic interpretations of text
- attempts to use evidence from text(s) to support claims; support may be limited or irrelevant

**0** A 0 point response shows little or no understanding of the passage main ideas and details.

## Scoring Criteria for Extended Response Reading Items

**ANALYSIS, INTERPRETATION, AND THINKING ABOUT TEXT:**

4 Points: Meets all relevant criteria

- thoroughly analyzes appropriate information and/or makes insightful connections between whole texts/parts of texts
- develops insightful interpretations of text
- uses ample, relevant evidence from text(s) to support claims

3 Points: Meets or most all relevant criteria

- analyzes appropriate information and/or makes thoughtful connections between whole texts/parts of texts
- develops thoughtful interpretations of text
- uses sufficient, relevant evidence from text(s) to support claims

2 Points: Meets some relevant criteria

- analyzes limited information and/or makes superficial connections between whole texts/parts of texts
- develops conventional or simplistic interpretations of text
- attempts to use evidence from text(s) to support claims; support may be limited or irrelevant

1 Point: Meets few relevant criteria

- shows difficulty analyzing information and/or makes weak connections between whole texts/parts of texts
- may not develop beyond literal interpretation of text
- uses little or no evidence to support claims

0 points - Student's response provides no evidence of interpretation or critical analysis of text required by the prompt; or the prompt may simply be recopied; or the response may be "I don't know" or a question mark (?).

## SCORING OF OPEN-ENDED MATHEMATICS ITEMS

Individual scoring criteria were developed for each constructed-response item. Short-answer items were scored on a scale of 0 to 2 points, and extended-response items were scored on a scale of 0 to 4 points. The following scoring criteria were used to guide item writers in their development of item specific scoring criteria. This helped to ensure that the item scoring criteria were clearly focused on the appropriate dimension of mathematics performance: conceptual and procedural understanding, mathematical problem-solving, mathematical communication, mathematical reasoning, OR mathematical connections.

## GENERAL SCORING CRITERIA FOR SHORT-ANSWER MATHEMATICS ITEMS

**MATHEMATICAL CONCEPTS AND PROCEDURES:**

**2** A 2-point response shows complete understanding of the concept or task, as well as consistent and correct use of applicable information and/or procedures. Set-up and computations are accurate.

**1** A 1-point response shows partial understanding of the concept or task. There may be minor errors in the use of applicable information and/or procedures. Set-up or computations may have minor errors.

**0** A 0 point response shows little or no understanding of the concept or task.

**COMMUNICATING MATHEMATICAL UNDERSTANDING:**

**2** A 2-point response shows understanding of how to effectively and appropriately interpret, organize, and/or represent mathematical information relevant to the concept.

**1** A 1-point response shows some understanding of how to interpret, organize, and/or represent mathematical information relevant to the concept; however, the response is not complete or effectively presented.

**0** A 0 point response shows little or no understanding of how to interpret, organize and/or represent mathematical information relevant to the concept.

**SOLVING MATHEMATICAL PROBLEMS:**

**2** A 2-point response shows thorough investigation, clear understanding of the problem, and/or effective and viable solution.

**1** A 1-point response shows partial investigation and/or understanding of the problem, and/or a partially complete or partially accurate solution.

**0** A 0-point response shows very little or no investigation and/or understanding of the problem, and/or no visible solution; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

## GENERAL SCORING CRITERIA FOR SHORT-ANSWER MATHEMATICS ITEMS (Cont.)

**MATHEMATICAL REASONING**

**2** A 2-point response shows effective reasoning through a complete analysis or thorough interpretation, supported predictions, and/or verification.

**1** A 1-point response shows somewhat flawed reasoning either through incomplete analysis or interpretation, prediction that lacks support, or inadequate verification.

**0** A 0-point response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?).

**MAKING MATHEMATICAL CONNECTIONS:**

**2** A 2-point response makes clear and effective connections within and/or between conceptual or procedural areas.

**1** A 1-point response makes vague or partially accurate connections within and/or between conceptual or procedural areas.

**0** A 0-point response makes little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or may indicate "I don't know" or a question mark (?)

## EXAMPLE OF SPECIFIC SCORING CRITERIA FOR A SHORT-ANSWER MATHEMATICS CONCEPTS AND PROCEDURES ITEM

Primary Essential Learning Requirement: Student ***understands and applies the concepts and procedures of mathematics: algebraic sense.***

Look at the following list of numbers. Describe **two** different patterns you see in these numbers.

9 18 27 36 45 54 63 72 81 90

**SCORING CRITERIA FOR ITEM**

**2** A 2-point response describes two different valid number patterns in the given list of numbers.

**1** A 1-point response describes only one valid number pattern in the list of numbers. Any alternate 1 point response describes two different number patterns but the descriptions may be vague, incomplete, or unclear.

**0** A 0 point response shows little or no understanding of number patterns.

# SCORING CRITERIA FOR EXTENDED-RESPONSE MATHEMATICS ITEMS

**SOLVING MATHEMATICAL PROBLEMS:**

4 points -- Meets all relevant criteria

- Thoroughly investigates the situation
- Uses all applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs elegant, efficient, valid solution using applicable tools and workable strategies

3 points -- Meets all or most relevant criteria

- Investigates the situation
- Uses most applicable information related to the problem
- Uses applicable mathematical concepts and procedures
- Constructs viable/acceptable solution using applicable tools and workable strategies

2 points -- Meets some relevant criteria

- Investigates the situation, but may omit issues or information
- Uses some applicable information related tot he problem
- Uses some applicable mathematical concepts and procedures
- Constructs solution using applicable tools and workable strategies, solution may not completely address all issues or strategies may have flaws

1 point -- Meets few relevant criteria

- Attempts to investigate the situation
- Uses some applicable information related to the problem
- Uses few applicable mathematical concepts and procedures
- Attempts solution, however, mostly incomplete or not effective

0-points--Student's response provides no evidence of problem-solving skills or shows very little or no understanding of the task; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## SCORING CRITERIA FOR EXTENDED-RESPONSE MATHEMATICS ITEMS (Cont.)

**COMMUNICATING MATHEMATICAL UNDERSTANDING:**

4 points -- Meets all relevant criteria

- Gathers all applicable information form appropriate sources
- Demonstrates interpretations and understandings in a clear, systematic, and organized manner
- Represents mathematical information and ideas in an effective format for the task, situation, and audience

3 points -- Meets most relevant criteria

- Gather applicable information from appropriate sources
- Demonstrates interpretations and understandings in a clear and organized manner
- Represents mathematical information and ideas in an expected format for the task, situation, and audience

2 points -- Meets some relevant criteria

- Gathers information from appropriate sources
- Demonstrates interpretation and understandings in an understandable manner
- Represents mathematical information in an acceptable format for the task, situation, and audiences

1 point -- Meets few relevant criteria

- Gathers little information from appropriate sources
- Demonstrates interpretations and understandings in a manner that may be disorganized or difficult to understand
- Represents mathematical information and ideas in a format that may be inappropriate for the task, situation, and audience.

0-points--Student's response shows little or no understanding of how to interpret, organize or represent mathematical information relevant to the concept; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## SCORING CRITERIA FOR EXTENDED-RESPONSE MATHEMATICS ITEMS (Cont.)

**MATHEMATICAL REASONING**

4 points -- Meets all relevant criteria

- Makes insightful interpretations, comparisons, or contrasts of information form sources
- Effectively uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes insightful conjectures and inferences, if asked
- Systematically and successfully evaluates effectiveness of procedures and results, if asked
- Gives comprehensive support for arguments and results

3 points -- Meets most relevant criteria

- Makes thoughtful interpretations, comparisons, or contrasts of information form sources
- Uses examples, models, facts, patterns, or relationships to validate and support reasoning.
- Makes expected conjectures and inferences, if asked
- Successfully evaluates effectiveness of procedures and results, if asked
- Gives substantial support for arguments and results

2 points -- Meets some relevant criteria

- Makes routine interpretations, comparisons, or contrasts of information form sources
- Includes examples, models, facts, patterns, or relationships to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Partially evaluates effectiveness of procedures and results, if asked
- Gives partial support for arguments and results

1 point -- Meets few relevant criteria

- Makes superficial interpretations, comparisons, or contrasts of information form sources
- Examples, models, facts, patterns, or relationships may not be included to validate and support reasoning.
- Conjectures and inferences, if given, may be naive
- Attends to wrong information and/or persists with faulty strategy when evaluating effectiveness of procedures and results
- Support for arguments and results may not be included

0-points--Student's response shows very little or no evidence of reasoning; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## SCORING CRITERIA FOR EXTENDED-RESPONSE MATHEMATICS ITEMS (Cont.)

**MAKING MATHEMATICAL CONNECTIONS:**

4 points -- Meets all relevant criteria

- Shows a thorough understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in a clear and insightful manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in a clear and insightful manner

3 points -- Meets most relevant criteria

- Shows a general understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines in an obvious/expected manner AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations in an obvious/expected manner

2 points -- Meets some relevant criteria

- Shows a partial understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in other disciplines AND/OR
- Identifies, analyzes, and/or applies mathematical patterns and concepts in real-life situations

1 point -- Meets few relevant criteria

- Shows a little understanding of links among areas of mathematics using equivalent representation AND/OR
- Identifies, mathematical patterns and concepts in other disciplines AND/OR
- Identifies applies mathematical patterns and concepts in real-life situations

0-points--Student's response makes very little or no connection within or between conceptual or procedural areas; or the prompt may simply be recopied, or the response may indicate "I don't know" or a question mark (?).

## SCORING OF WRITING ITEMS

Students write in response to two prompts. Scoring criteria for two traits are applied to each piece of writing. One trait includes scoring for content, style, and organization of the writing, and the other trait includes scoring for the writing conventions (spelling, capitalization, punctuation, grammar). These scoring criteria are not adapted to the specific demands of a writing prompt since students have many choices about the topics for their writing and for the ways in which they apply stylistic elements.

**CONTENT, STYLE, AND ORGANIZATION**

4 points

- Maintains consistent focus on the topic and has ample supporting details
- Has logical organizational pattern and conveys a sense of wholeness and completeness
- Provides transitions which clearly serve to connect ideas
- Uses language effectively by exhibiting word choices that are engaging and appropriate for the intended audience and purpose
- Includes sentences of varied length and structure
- Allows the reader to sense the person behind the words

3 points

- Maintains adequate focus on the topic and has adequate supporting details
- Has logical organizational pattern and conveys a sense of wholeness and completeness, although some lapses may occur
- Provides adequate transitions in an attempt to connect ideas
- Uses effective language and appropriate word choices for the intended audience and purpose
- Includes sentences that are somewhat varied in length and structure
- Provides the reader with some sense of the person behind the words

2 points

- Demonstrates an awareness of the topic and includes some (or few) supporting details, but may include extraneous or loosely related material
- Shows an attempt at an organizational pattern, but exhibits little sense of wholeness and completeness
- Provides transitions that are weak and inconsistent
- Has a limited and predictable vocabulary that may not be appropriate for the intended audience and purpose
- Shows little variety in sentence length and structure
- Attempts to give the reader a sense of the person behind the words

1 point

- Presents minimal information or ideas and few supporting details which may be inconsistent or interfere with the meaning of the text
- Has little evidence of an organizational pattern or any sense of wholeness and completeness
- Provides transitions that are poorly utilized or fails to provide transitions
- Has a limited or inappropriate vocabulary for the intended audience and purpose
- Has little or no variety in sentence length and structure
- Provides the reader with little or no sense of the person behind the words

0-points--response is "I don't know"; response is a question mark (?); response is one word; response is only the title of the prompt; or the prompt is simply recopied.

## SCORING OF WRITING ITEMS (Cont.)

**CONTENT, STYLE, AND ORGANIZATION**

2 points

- Consistently follows the rules of standard English for usage, spelling of commonly used words, capitalization, punctuation, and sentence formation
- Exhibits the use of complete and fluent sentences except where purposeful phrases or clauses are used for effect
- Indicates paragraphs consistently

1 point

- Fairly consistently follows the rules of standard English for usage, spelling of commonly used words, capitalization, punctuation, and sentence formation
- Generally exhibits the use of complete and fluent sentences except where purposeful phrases or clauses are used for effect
- Indicates paragraphs for the most part

0 points

- Basically does not follow the rules of standard English for usage, spelling of commonly used words, capitalization, punctuation, and sentence formation, although some elements may be correct
- Exhibits errors in sentence structure that impede communication
- Indicates paragraphs only to a limited degree

OR

- Response is "I don't know", a question mark, one word, only the title of the prompt, or prompt is simply recopied.

# APPENDIX E

## WASHINGTON ASSESSMENT OF STUDENT LEARNING

### National Technical Advisory Committee

### Washington State Assessment Advisory Team

**National Technical Advisory Committee Members**

Peter Behuniak, Director of Testing, Connecticut State Department of Education

Robert Linn, Professor, University of Colorado and UCLA/CRESST

William Mehrens, Professor, Michigan State University

Joseph Ryan, Professor, Arizona State University

Kenneth Sirotnik, Professor, University of Washington

**Washington State Assessment Advisory Team**

Gordon Ensign, Director of Assessment, Commission on Student Learning

Kathy Kimball, Assistant Director, Commission on Student Learning

Duncan MacQuarrie, Director of Research and Evaluation, Washington State Office of the Superintendent of Public Instruction

Geoff Praeger, Director of Testing, Central Valley School District

Nancy Skerritt, Director of Curriculum, Tahoma School District

Catherine Taylor, Associate Professor, University of Washington

Joe Willhoft, Director of Research and Evaluation, Tacoma School District